THE UNIVERSITY *of* EDINBURGH

Computational Approaches for Identifying Inhibitors of Protein Interactions

Thesis Submitted for the Degree of

Doctor of Philosophy

Wissam Mehio BSc. MSc.

Structural Biochemistry Group

The Institute of Structural and Molecular Biology

University of Edinburgh

October 2010

Abstract

Inter-molecular interaction is at the heart of biological function. Proteins can interact with ligands, peptides, small molecules, and other proteins to serve their structural or functional purpose. With advances in combinatorial chemistry and the development of high throughput binding assays, the available inter-molecular interaction data is increasing exponentially. As the space of testable compounds increases, the complexity and cost of finding a suitable inhibitor for a protein interaction increases. Computational drug discovery plays an important role in minimizing the time and cost needed to study the space of testable compounds. This work focuses on the usage of various computational methods in identifying protein interaction inhibitors and demonstrates the ability of computational drug discovery to contribute to the ever growing field of molecular interaction.

A program to predict the location of binding surfaces on proteins, STP (Mehio et al., Bioinformatics, 2010, *in press*), has been created based on calculating the propensity of triplet-patterns of surface protein atoms that occur in binding sites. The use of STP in predicting ligand binding sites, allosteric binding sites, enzyme classification numbers, and binding details in multi-unit complexes is demonstrated. STP has been integrated into the in-house high throughput drug discovery pipeline, allowing the identification of inhibitors for proteins whose binding sites are unknown.

Another computational paradigm is introduced, creating a virtual library of β-turn peptidomimetics, designed to mimic the interaction of the Baff-Receptor (Baff-R)

with the B-Lymphocyte Stimulator (Blys). LIDAEUS (Taylor, et al., Br J Pharmacol, 2008; 153, p. S55-S67) is used to identify chemical groups with favorable binding to Blys. Natural and non-natural sidechains are then used to create a library of synthesizable cyclic hexapeptides that would mimic the Blys:Baff-R interaction.

Finally, this work demonstrates the usage and synergy of various in-house computational resources in drug discovery. The ProPep database is a repository used to study trends, motifs, residue pairing frequencies, and aminoacid enrichment propensities in protein-peptide interaction. The LHRLL protein-peptide interaction motif is identified and used with UFSRAT (S. Shave, PhD Thesis, University of Edinburgh, 2010) to conduct ligand-based virtual screening and generate a list of possible antagonists from the EDULISS (K. Hsin, PhD Thesis, University of Edinburgh, 2010) compound repository. A high throughput version of AutoDock (Morris, et al., J Comput Chem, 1998; 19, p. 1639-62) was adapted and used for precision virtual screening of these molecules, resulting in a list of compounds that are likely to inhibit the binding of this motif to several Nuclear Receptors.

Declaration

The work presented in this thesis is the original work of the author. This thesis has been composed by the author and has not been submitted in whole or in part for any other degree.

Wissam Mehio

Acknowledgements

I would like to thank my supervisors Prof Malcolm Walkinshaw and Dr Paul Taylor for the stimulating advice and guidance they have provided during this work.

I also thank The Darwin Trust of Edinburgh for financial support; Dr Graham Kemp for his help with the STP algorithm; Dr Janice Bramham for her help with the complement immune system; Prof Manfred Auer and Dr Martin Hintersteiner for their help with the Blys/BaffR study. I thank the members of the Walkinshaw group, especially Dr Hugh Morgan for his help with the study of Pyruvate Kinase; Dr Matthew Nowicki for his help with CRK3; Dr Douglas Houston for his help with the automated virtual screening pipeline; Dr Iain McNae for his help with cyclophilin; and Dr Simon Harding for his help with the ProPep database. Many thanks go to Dr Kun-Yi Hsin and Dr Steven Shave for their input and help with all the computational tasks created and discussed within this work.

My deepest gratitude goes to Mr Mohammad al Amin Itani and Mr Khodor Alameh for their invaluable support leading to the commencement and completion of this work.

Last but not least, special thanks go to Oussama, Fadwa, Carole, Maher, Hadi, and Majd for a lifetime's worth of support without which, this work would not have been completed.

Table of Abbreviations

3D	Three-dimensional
ABL	Abelson Leukemia virus
ACS	American Chemical Society
ADMet	Absorption, distribution, metabolism, excretion, and toxicity
ADP	Adenosine-diphosphate
Ala	Alanine
aLogp	Ghose-Crippen octanol-water partition coeff
APRIL	A proliferation inducing ligand
Arg	Arginine
Asn	Asparagine
Asp	Aspartic Acid
ATP	Adenosine-5'-triphosphate
AVG	Average
BAFF	B cell-activating factor
BCMA	B Cell Maturation
BLAST	Basic local alignment search tool
BLOSUM	Blocks of amino acid substitution matrix
BLys	B lymphocyte stimulator
CAS	Chemical Abstracts Service
CATH	Class, Architecture, Topology, Homologous superfamily
CBF3	Centromere Binding Factor 3
CDK	Cyclin dependent kinase
CDP	Cell division control protein 4 – phosphodegron
CDS	Chemical Database Service
Cha	Cyclo Hexyl Alanine
CKS1	Cdc kinase subunit 1
Cpa	Cyclo Pentyl Alanine
CPD	Cdc4-phosphodegron
CRD	Cys-rich domains
CRK	Cyclin related kinase

CUL1	Cullin 1
Cys	Cysteine
DADH	D-Arg Dehydrogenase
Dhr	D-Homo Arginine
Dlu	D-Luecine
Dmb	Di-amino Butanoic
DNA	Deoxyribonucleic acid
Drg	D-Arginine
DS Visualiser	Discovery Studio Visualiser
EC Numbers	Enzyme Comission numbers
ECP	Enzyme class propensity
EDULISS	Edinburgh University ligand selection system
ELM	Eukaryotic Linear Motif
EM	Electron Microscopy
F-1,6-BP	Fructose 1,6-bisphosphate
F-2,6-BP	Fructose 2,6-bisphosphate
Fg2	2-Fluoro Phenyl Gly
Fg3	3-Fluoro Phenyl Gly
Fg4	4-Fluoro Phenyl Gly
Fgl	Phenyl Gly
FKPB	FK506 binding protein
FSH	Follicle Stimulating Hormone
Gln	Glutamine
Glu	Glutamic Acid
Gly	Glycine
GnRH-R	Gonadotropin-releasing hormone receptor
GPCR	G protein-coupled receptor
GPU-CUDA	Graphics processing unit- compute unified device architecture
GWIDD	Genome-wide protein docking database
HB	Hydrogen Bond
Hfl	Hexa Flouro Leucine
Hgu	Homo Glutamic Acid

His	Histidine
HIV	Human immunodeficiency virus
Hrg	Homo Arginine
HSD	Hydroxysteroid dehydrogenase
HSQC	Heteronuclear Single Quantum Coherence
Htr	Homo Threonine
HTS	High throughput screening
IC-50	Half maximal inhibitory concentration
ICB	Integrated Chemical Biophysics
Ile	Isoleucine
Kd	Dissociation constant
Ki	Absolute inhibition constant
LBD	Ligand binding domain
Leu	Leucine
LH	Luteinising Hormone
LIDAEUS	Ligand Discovery at Edinburgh University
Lys	Lysine
MAC	Membrane Attack Complex
Mcf	M-Carboxy-Phenyl Alanine
MDM2	Murine double minute 2
Met	Methionine
MHC-II	Major Histocompatibility complex class II
MISCC	Maybridge, InterBioScreen, Specs, Chemdiv, and Chembridge
MM Energy	Molecular Mechanics Energy
MM Minimization	Molecular Mechanics Minimization
mRNA	Messenger ribonucleic acid
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NMR	Nuclear magnetic resonance
Ocf	O-Carboxy-Phenyl Alanine
ORF	Open Reading Frame
PAM	Point accepted mutation

Pcf	P-Carboxy-Phenyl Alanine
PCNA	Proliferating Cell Nuclear Antigen
Pda	Piperidine Alanine
PDB	Protein Data Bank
PEP	Phosphoenolpyruvate
Phe	Phenylalanine
PLB	Propensity for ligand binding
Ppg	Phenyl Propyl Glycine
PPIases	Peptidylprolylisomerase
Pro	Proline
PXR	Pregnane X receptor
PyK	Pyruvate Kinase
Pza	Piperazinyl Alanine
QT Clustering	Quality- Threshold clustering
RASA	Reduction in accessible surface area
RMS	Root mean square
RMSD	Root mean square deviation
RNA	Ribonucleic acid
RPP	Residue Pairing Prefrence
SAR	Structure-activity relationship
SCFCDC4	SKP1-CUL1-F-box protein - cell division control protein 4
SCOP	Structural Classification of Proteins
Ser	Serine
SH3	Sarcoma Homology 3
SKP1	S-phase kinase-associated protein 1
SKP2	S-phase kinase-associated protein 2
SNP	Single Nucleotide Polymorphism
SpeA	Streptococcal pyrogenic exotoxin A
Stdev, SD	Standard Deviation
STP	Surface triplet propensities
STPWater	Surface triplet propensities with water
SVM	Support vector machine

TACI	Transmembrane activator and CAML interacting protein
Tbu	Tetra Butanoic
TCP	Triplet class propensity
TCR	T-Cell Receptor
Tfa	Thio Phenyl Alanine
Tfg	Thio Phenyl Glycine
Tfl	Tri Flouro Leucine
Thr	Threonine
TNF	Tumor necrosis factor
TNFR	Tumor necrosis factor receptor
TRAF	TNF Receptor Associated Factor
tRNA	Transfer ribonucleic acid
Trp	Tryptophan
Ttz	Tetrazole
Tyr	Tyrosine
Tza	Tetrazole Alanine
UCSF	Unviersity of California at San Francisco
UFSRAT	Ultra fast shape recognition with atom types
Val	Valine
VDW	van der Waals
XIAP	X chromosome-encoded inhibitor-of-apoptosis protein
ΔG_{stat}	Statistical free energy

Table of Contents

1 Computational Methods for Molecular Recognition

1.1 Molecular Function

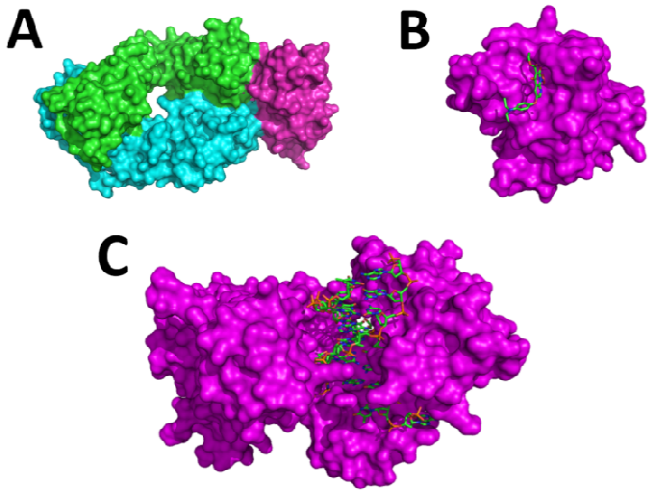

Figure 1-1: The roles played by proteins in various molecular functions. A shows the binding of the HyHel Antibody (blue and green) to the Lysozyme antigen (PDB structure 3HFM [1]). B shows the binding of an inhibitor molecule to the binding site of HIV-1 Reverse Transcriptase, inhibiting its RNaseH activity (PDB structure 3LP3 [2]). C shows the Eta Human DNA polymerase in complex with the DNA (PDB structure 3MR2 [3]).

Biological function is a result of molecular interaction. Proteins constitute the main functional and structural building blocks in organisms. Digestion, immunity, motion, metabolism, cell growth and signaling are all controlled by protein interaction. Proteins interact with themselves to make up larger functional complexes, and they also interact with small molecules like peptides, sugars, and other ligands. These interactions are also vital to the molecular functions and have proved to be the backbone of drug design theory (Figure 1-1). Consequently, understanding these interactions can shed some light on different biochemical pathways in organisms,

and can help scientists use this new information to uncover the mysteries that lie within organisms.

Research on intermolecular interactions was originally dependent on wet lab experiments. Unfortunately, these experiments require a lot of time, effort, skill and money. With the emergence of powerful computing capabilities, and as computers have become cheap and integrated in everyday life, the *in-silico* research paradigm has slowly started proving itself to be of great importance [4-13]. The ultra fast computer speed and the low cost per experiment allow scientists to repeat the same experiments over and over again at little additional cost.

1.2 Computational Biology

Computational biology plays an important role in the study of molecular interaction. The continuous increase in computational power [14] has made it possible to both simulate previously unfeasible large and complex systems and at increase the level of accuracy. For example, a docking experiment with modern GPU-CUDA technology [15] is simulated in less than 5 minutes [16]. Twenty years ago, the same experiment would have taken days or even weeks. This enormous speed makes computational biology a useful partner for wet lab experiments. The applications of computational biology are vast and widespread. These applications include tasks like simple data storage, genomic searching, protein modeling, protein structure prediction, protein interaction prediction, and drug discovery. The contribution of computational methods to each of these areas is discussed below.

1.2.1 Electronic Databases

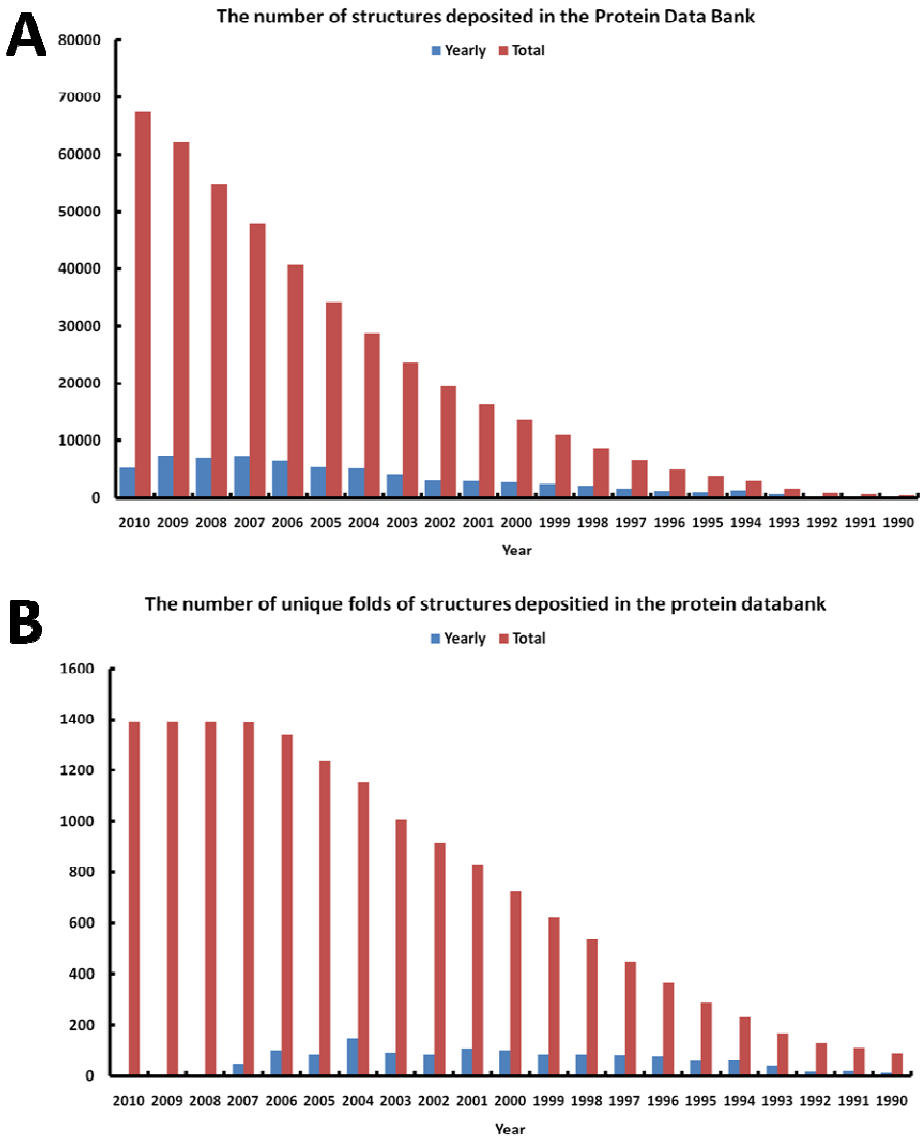

Figure 1-2: The increase of deposited structures in the Protein Data Bank since 1990. Numbers from 1972 to 1990 are small and hence not shown. Number of yearly deposited structures was steadily increasing between 1990 and 2006, and appears to have reached a plateau starting 2007. Nevertheless, the PDB now houses over 67,000 structures and this provides a rich resource of information for computational studies. Interestingly, no new folds have been discovered since 2007.

The simplest form of participation of the *in silico* methods in studying molecular interaction is the electronic storage of large quantities of data and subsequently

publishing these databases online along with fast searching tools. The Protein Data Bank (PDB) [17] houses more than 67000 protein coordinate structures (58.5K Xray, 8.5K NMR, 302 EM). This database has been continuously growing (Figure 1-2), providing a comprehensive resource of three dimensional coordinates of protein structures, covering 1393 SCOP folds [18] and 1233 CATH topologies [19, 20]. This resource is at the heart of all the statistical analyses computed on protein structures, folds, and interactions.

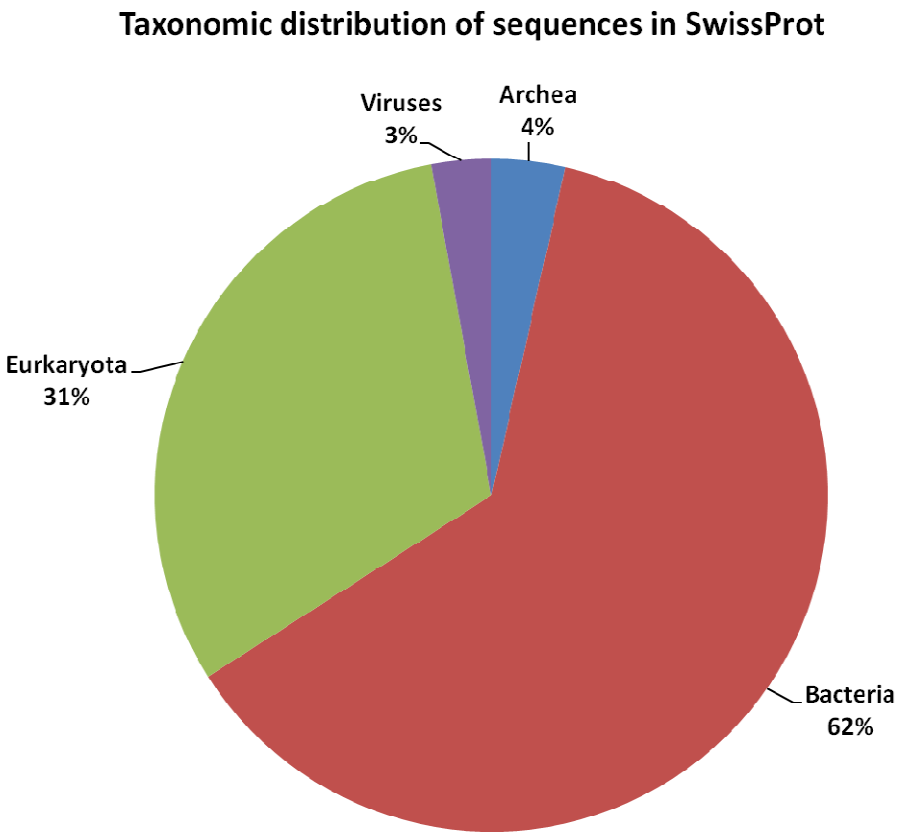

Figure 1-3: The distribution of Swissprot protein sequences across the 4 kingdoms of organisms.

The UniProt Knowledgebase/SwissProt [21] is another widely used database, holding more than 500,000 annotated protein sequences from Archea, Bacteria, Eukaryota, and Viruses (Figure 1-3). This database is equipped with various mining programs like the basic local alignment search tool (BLAST) [22] for similarity searches, MotifScan [23] for pattern searching, and T-Coffee [24] and ClustalW [25] for sequence alignment. The incorporation of all these tools within a single web interface has provided researchers with means of conducting fast computational data mining operations.

Another important computational contribution to the field of molecular interaction is that of the NCBI at NIH [26], through the databases like PubMed[1] and PubChem[2]. With Pubmed, searching all biological and medical journal papers is possible with several advanced search options to filter by date, author, journal, and keywords. Pubchem is a similar resource, dealing with all published chemical compounds, with bioactivity annotation and links to the bioassays used in testing them. The NIH also includes specialist databases for Nucleotide[3], Single Nucleotide Polymorphism[4] (SNP), and many more, designed for conducting detailed searches of published nucleotide sequences and SNPs.

These databases and others provide an instantaneous and free means of searching all the documented knowledge in the field of molecular interaction. They play an

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://pubchem.ncbi.nlm.nih.gov/
[3] http://www.ncbi.nlm.nih.gov/nucleotide/
[4] http://www.ncbi.nlm.nih.gov/snp/

important role in the unification of data and the developments of common standards needed to facilitate the storage, annotation, and mining of all the data produced by the worldwide scientific community.

1.2.2 Gene homology and Protein Sequence

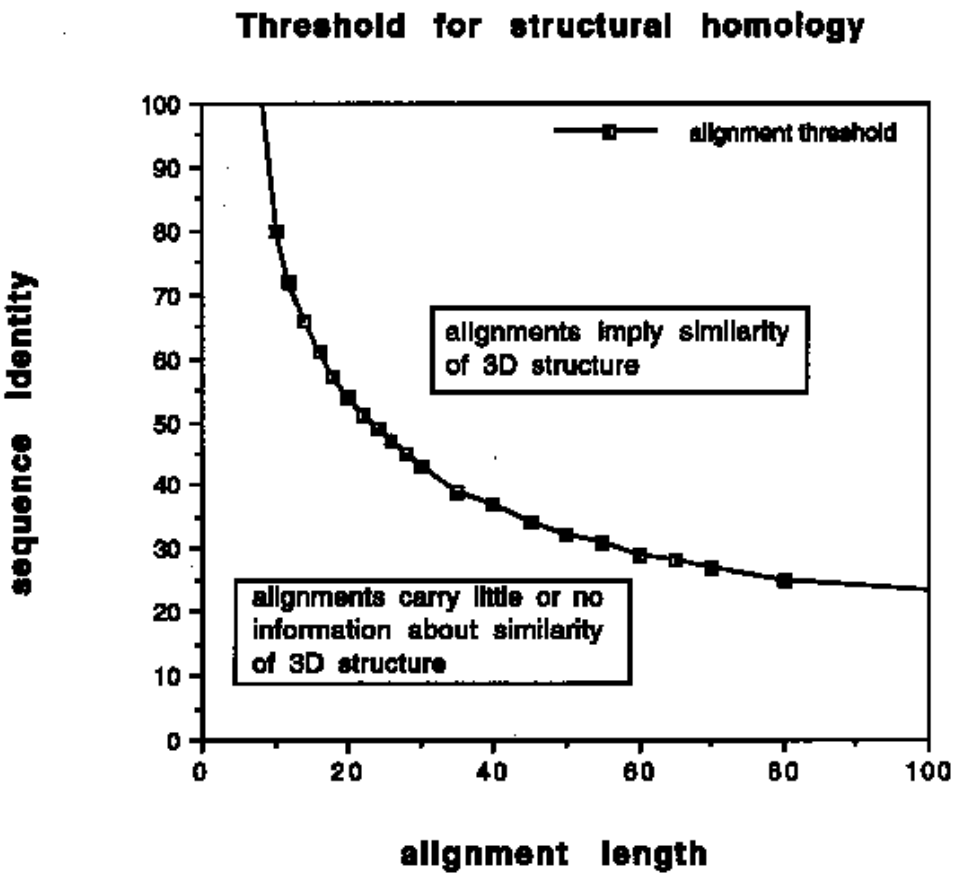

Figure 1-4: The relationship between sequence identity and three dimensional similarity: a minimum of 30% sequence identity is needed to infer a structural similarity. Research done by Sander and Schneider [27], Figure adapted from Rodriguez and Vriend [28].

Gene expression provides large room for the input of computational biology especially given the major progress of informatics in other areas of pattern matching (regular expressions in text [29] and language syntax definitions [30]). This problem

could be applied to the DNA searching routines, where the genome is read as long sequences of nucleotides, and then pattern matching algorithms could find the ORFs in that genome. Another application is centered on an observation made by Sander and Schneider [27] (Figure 1-4), who linked sequence similarity with protein function. This led to the development of tools like Blast [22], T-Coffee [24], and ClustalW [25] that would compute the sequence similarity between two or more DNA/Protein sequences. These programs are now highly evolved and use mutation matrices like BLOSUM (blocks of amino acid substitution matrix) [31] and PAM (Point Accepted Mutation) [32] substitution matrices, which give mutation scores between amino acids (rather than using the original hit or miss paradigm).

1.2.3 Protein fold and structure

Proteins are described according to four structural levels . The primary structure of a protein is the sequence of aminoacids that constitute that protein. The secondary structure of a protein classifies stretches of aminoacids according to their geometric configuration: α-helices and β-sheets. The tertiary structure of a protein is the final 3-dimensional conformation that the entire protein takes, by folding these helices, sheets, and loops to form a specific structure. The quaternary structure describes the arrangement in which the different chains and subunits of a certain protein complex undergo to form the larger complex. This sequence and geometric description of the protein structure gives rise to several applications like structural alignment and fold categorization.

Sequence similarity usually leads to structural similarity. However, structural similarity among proteins is more conserved than sequence similarity, with remote

protein homologues having dissimilar sequences while their 3D structures are highly conserved [33, 34]. The motivation behind structural similarity of proteins is that two proteins with a similar structure are expected to bind similar (or the same) molecules [35]. Thus, the need for the structural alignment of protein domains presents itself, especially with the increase of the number of structurally determined proteins (Figure 1-2). Many structural similarity programs are currently available (for eg, FAST [36], MultiProt [37], and SuperPose [38]). Algorithms in this field vary between matching-based techniques (3D sequence-independent structural comparison [39, 40]), optimization based techniques (genetic algorithm-based protein structure comparisons [41], and grid-based techniques [42]. Currently, these structural superposition algorithms have found their way to be subroutines of molecular imaging software like Pymol [43], UCSF Chimera [44], and Accelrys DS Visualiser [45].

Structural and fold classification of protein domains is a step forward from structural similarity algorithms, aimed at combining primary, secondary and tertiary structure information to cluster proteins into structural and functional families. Perhaps, the most famous methods of protein fold classification are CATH [19, 20] and SCOP [18] (the primary methods used in the PDB to predict protein folds). CATH is a hierarchic classification of protein domains based on protein class (C), architecture (A), topology (T), and homologous superfamily (H). Class refers to the secondary structure of a domain, architecture refers to the overall shape of the orientations of these secondary structures (manually assigned), topology refers to the similarity of secondary structure at the domain core, and homologous superfamily corresponds to

high sequence and structural similarity between the domains [19]. SCOP (Structural Classification of Proteins) is based on the evolutionary and structural relationships of all structurally known proteins (classification is also per domain like CATH). These classification paradigms demonstrate yet another contribution of computational theory to the field of molecular recognition, helping with summarizing complex geometric shapes into words and quantifiable substructures and providing more insight into the prediction of the function of protein domains from their primary, secondary, and tertiary structural descriptions.

1.2.4 Predicting Protein function

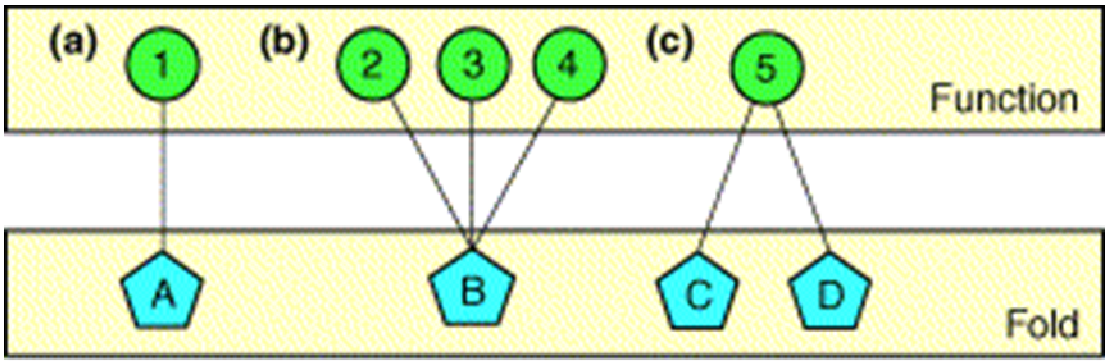

Figure 1-5: The relationship between protein fold and function is not a strictly one to one relationship (a). Proteins with similar folds can have different functions (b), leading to a one to many relationships. Many to one relationships also exist (c), where proteins with different folds exhibit the same function [46].

The relationship between the fold the function is not always straightforward (Figure 1-5). Proteins with the same folds, for example TIM barrels, can have multiple functions [47] while proteins with different folds share the same function [48]. In

fact, a more accurate approach to predict the function of a protein is through the study of the binding site of that protein. Some protein function prediction methods rely on the geometry of a binding site as similar geometries would bind similar ligands and hence, perform a similar function [49]. Geometric hashing [50] is one of the widely used paradigms for the geometric comparison of protein surfaces, utilizing the same methodology behind computational face recognition to identify topological patterns of protein surfaces. Other algorithms utilize physiochemical annotation as well as geometric patterns to define pocket similarities. Examples of these physiochemical properties include hydrophobicity and electrostatic potential [51], amino acid distributions [52], and atom type classifications [48].

Modern function prediction algorithms are classified into several groups [53]. Similarity group methods [54-56] utilize sequence similarity searches against large databases of known proteins. Phylogenomic approaches [57, 58] build phylogenetic trees of all evolutionary homologues of a certain protein based on sequence similarity. Pattern based methods [59-61] make use of locally conserved sequence patterns instead of global sequences. Clustering approaches [62, 63] try to cluster all the known sequences, leading to the functional characterization of the un-annotated sequences. Finally, machine learning methods [64-67] utilize artificial intelligence and to learn characteristic combinations of protein properties and use them in the prediction of protein functions.

1.2.5 Docking

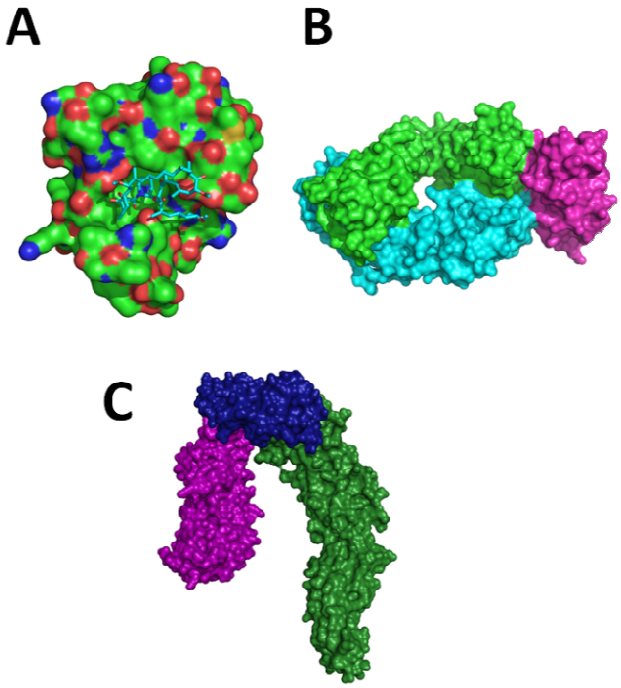

Figure 1-6: The three different docking categories. (A) shows the binding of a small molecule (FK506) to FKB12. (B) shows lysozyme (magenta) binding to the HyHel Fab antibody. (C) shows the multiprotein binding of SKP1 (blue), SKP2 (magenta), and CUL1 (green) to form the E3 Ligase.

Docking algorithms are another important contribution by informatics to the field of molecular recognition. These programs try to simulate the interaction between molecules by binding, or docking, a target molecule to a receptor molecule, and scoring that conformation using an energy function. This docking problem however may be classified into 3 categories. The first category consists of simulations that try to dock small molecules onto a certain protein (Figure 1-6 A). The second category consists of simulations that try to dock pairs of protein molecules (Figure 1-6 B). The third category is the most complex, and consists of simulations to dock several molecules together to form large multimolecular assemblies (Figure 1-6 C).

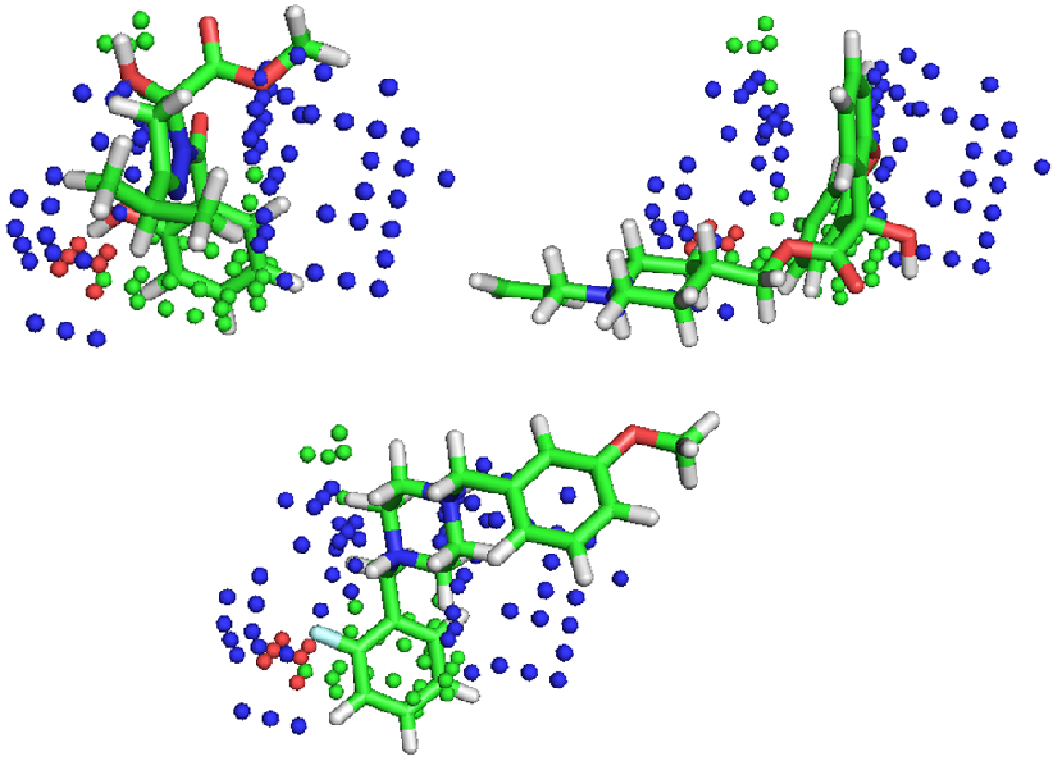

Figure 1-7: The fitting of several ligands (in sticks) onto the site points (points) generated by LIDAEUS to perform a fast docking simulation [68].

The first category of docking tasks is perhaps the simplest (yet not simple at all). With one of the interacting partners being a small molecule (referred to as the ligand), the amount of flexibility on that ligand is limited in comparison with the flexibility of the receptor protein. This decreases the search space of the docking algorithm. Even so, the space of conformations is still huge and limiting it is always a preferable option prior to beginning the docking simulations (for eg, by confining the searchable protein surface to a limited patch where the interaction is thought to occur). Our group currently relies on two protein-ligand docking algorithms. LIDAEUS [69] is a fast docking algorithm, capable of screening more than 4 million compounds in around 14 hours. It relies on limiting the searchable protein surface to a small patch, and creates site points of interaction surrounding that interaction patch. Then ligands are quickly fitted onto these site points via rigid body docking (Figure

1-7) and the conformations are scored and ranked using parameters similar to the Tripos forcefield [70]. AutoDock [71] performs more accurate searches by simulating ligand bond rotamers for the ligand. It is also possible to include receptor side chain flexibility in the simulations as well. As a result of this refinement, AutoDock simulations take around 5 minutes per ligand (which makes screening a 4-million -compounds-library a near impossible task). Our normal procedure is to prescreen a large dataset with LIDAEUS and then re-dock the top N hits with AutoDock to produce high resolution results.

The second category of docking involves the docking of two large molecules. This type of docking is computationally expensive, since the rotation and translation of every molecule is to be considered, resulting in a 6-dimensional search. Unlike the protein – small molecule docking, each rotation or translation of any of the partner molecules results affects a large number of atoms. The energy calculations also involve a very large number of close atoms (as the protein-protein interface is usually large). This increases the complexity of protein-protein docking simulations. Many algorithms are available to simulate this type of docking like Hex [72], RosettaDock [73], and FibreDock [74]. Hex uses *spherical polar Fourier* correlations to define the 3D shape of the docking partners and perform shape complimentarity calculations. Docking refinements with molecular mechanics and electrostatic interactions can also be used. RosettaDock searches the rigid body and side chain conformational space of two interacting partners to find a minimum energy complex using Monte-Carlo minimization simulations [75]. FibreDock is an advancement of the FireDock [76] program. It uses normal modes [77-79] to simulate backbone flexibility and

integer linear programming [80] to optimize side chain conformations. Monte-Carlo simulations are then used to calculate the energy of interaction and hence rank the different docking orientations.

The third category of docking simulations involves docking multiple molecules together to create a large multimolecular complex. Three of the methods that tackle this task are CombDock [81], EMatch [82], and MultiFit [83]. CombDock requires high (atomic) resolution structures and performs the docking task by docking all pairs together then heuristically predict the overall final assembly. EMatch deals with intermediate resolution cryo-EM density map of the complex, and fits the individual structures onto this map to create the multimolecular complex. MultiFit accepts low resolution density maps of the assembly and atomic resolution structures, and tries to fit those structures together in a puzzle solving technique.

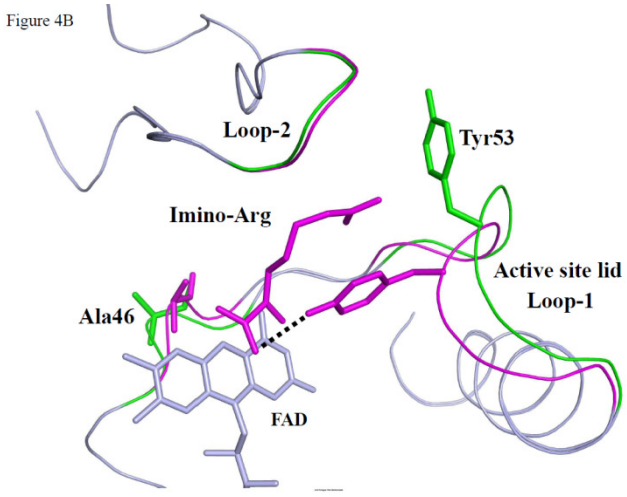

Figure 1-8: Comparison of ligand-free conformation (green) and product-bound conformation (magenta) at loop L1 (residues 33-56) and L2 (residues 244-248) regions in DADH structure shows the conformational shift that these loops undergo [84].

All docking simulations face the challenges of protein flexibility, induced fits, allosteric interactions, and entropy calculations as these events and calculations are often unpredictable. The three dimensional structures we have for the proteins are snapshots describing the state of these constantly moving molecules at a certain point of time. While parts of the protein stay rigid or exhibit minimal change of conformation, side chains (especially long ones like Arg) and loops are often mobile (for example, the displacement of the binding site loops in Pseudomonas aeruginosa D-Arg Dehydrogenase [84], Figure 1-8). In other cases, the introduction of a certain compound might lead to a change in the conformation of the protein atoms surrounding it, leading to an induced fit of that molecule into the pocket of the receptor (for example, the binding of bisphenol A and 4-α-cumylphenol by ERRGamma [85], Figure 1-9). In most cases, this induced fit is necessary for the binding of the ligand to the receptor, and trying to assess the interaction between the *apo* structure of the protein and that ligand leads to a weaker binding assessment. Allosteric interactions are a special form of induced fits, where the conformational change induced by a ligand binding to a protein takes place at a remote location on that protein, creating new binding sites and/or enhancing current ones (e.g. the allosteric alterations in pyruvate kinase upon binding to fructose 2,6 bisphosphate [86] (Figure 1-10). Docking programs cannot predict these allosteric changes and thus fail to accurately characterize the interaction of the ligand. Finally, the last hurdle for docking programs relates to the fundamental calculations of the binding energy of two molecules, which is divided into enthalpy and entropy. Enthalpy corresponds to the covalent, van der Waals, hydrogen bond, and electrostatic energies and these are well characterized. Entropy, on the other hand, denotes the

energy that is needed to move the system from a state of disorder (or chaos), to the ordered state in which the binding has occurred. This is a difficult calculation including aspects of global solvent calculations and approximation of contributions from side chain mobility. In the absence of a fast and accurate method of entropy estimation it is hard to assess the exact binding energy between interacting partners.

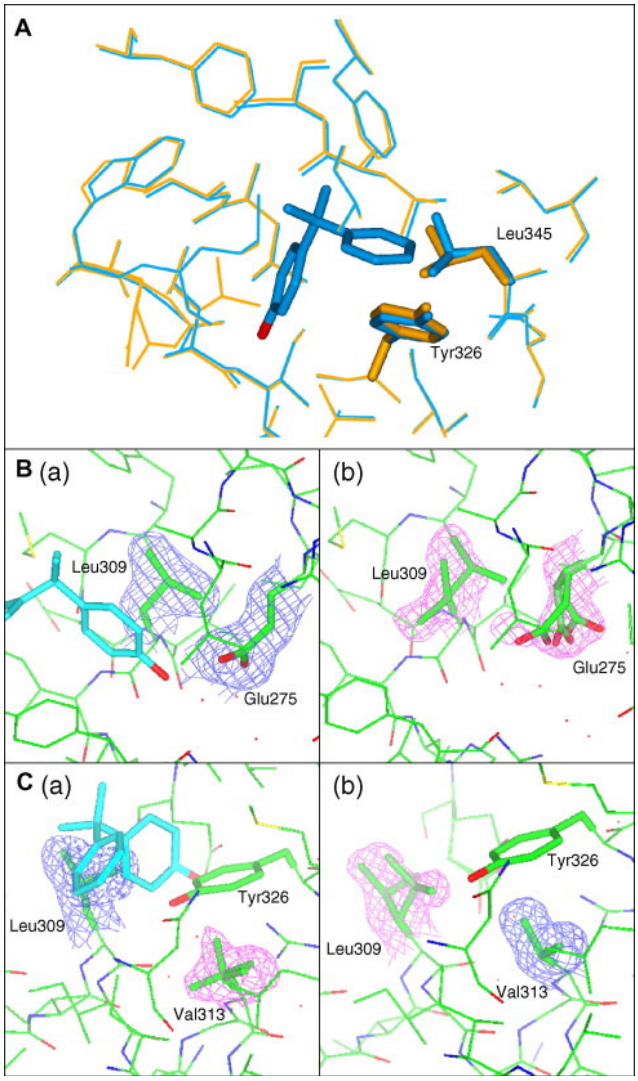

Figure 1-9: Induced-fit binding of 4-α-cumylphenol to the ERRγ-LBD apo form. (A) shows the superposition of residues in close proximity to the ligand (orange for apo conformation, and blue for induced conformation). (B-a) shows induced re-positioning of Glu275 and Leu309 by the binding of 4-α-cumylphenol and (B-b) their conformation in the apo form. (C-a) shows Val313 in its induced conformation while (C, b) shows the apo form [85].

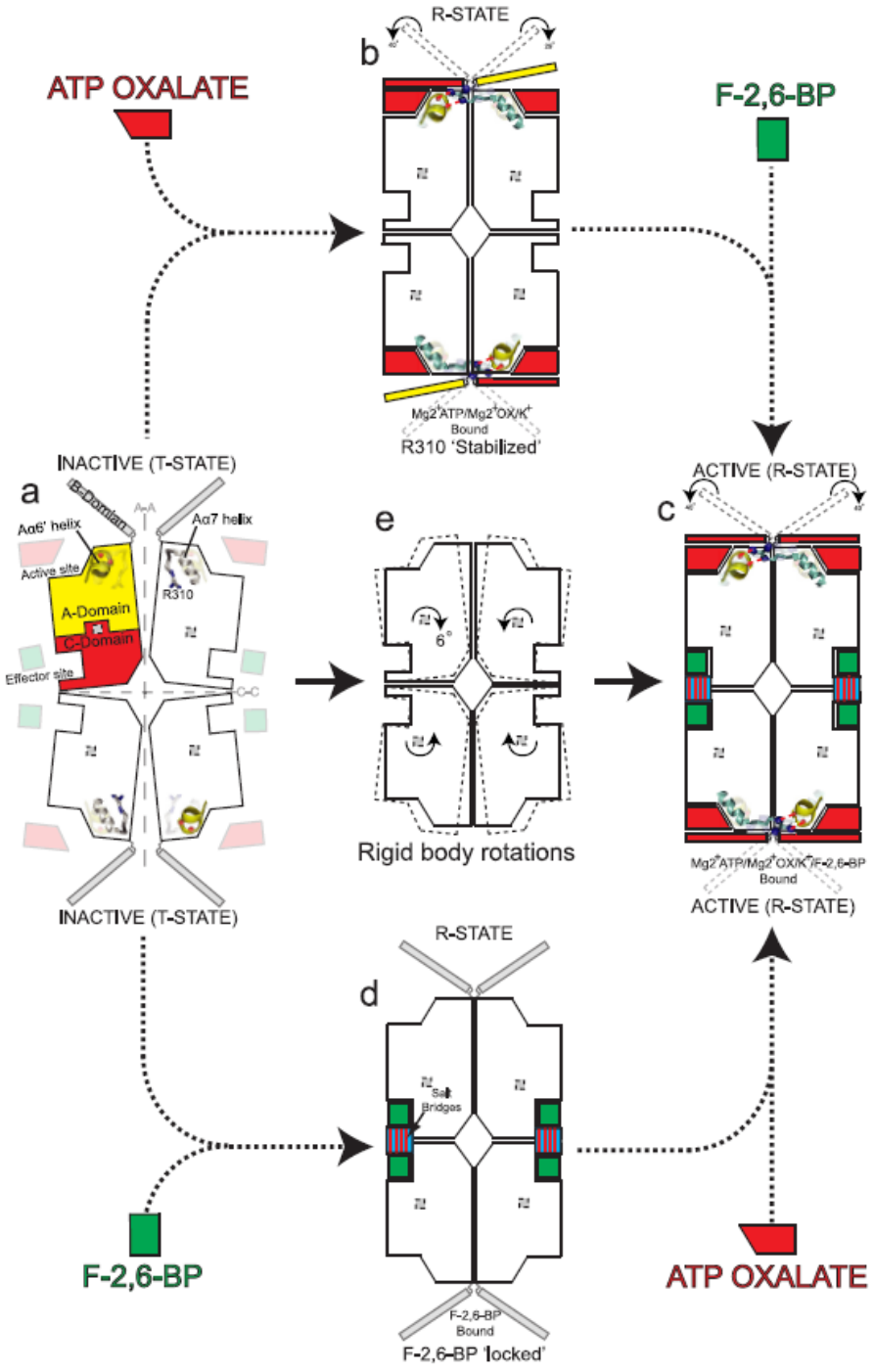

Figure 1-10: The transition of PyK between the inactive T state and the active R state, with the help of the allosteric ligand F-2,6-BP to stabilize the structure in the R-state. Picture taken from Morgan et al., (2010) [86].

However, these challenges do not render docking simulations useless. Modern docking programs have become more successful at realistically simulating binding interactions and identifying ligands that bind to the receptors *in vitro*. Moreover, the monetary and time advantages possessed by these algorithms are huge. LIDAEUS [69] docks 4 million small molecules onto a receptor within 14 hours. Autodock [71], which is a more accurate and thus, "slow", docks a small molecule within 5 to 10 minutes. Hex [72] currently docks 2 large globular proteins in less than 5 minutes by using the new GPU-CUDA technology [15]. This high speed combined with the fact that these simulations can be run at negligible cost (compared to the costs of buying materials and running complex screening robots in wet lab) make docking a very desirable option and one of the most important *in silico* contributions to the field of molecular recognition.

1.3 Computational Drug Discovery and Virtual Screening

Drug discovery is an important application of the study of molecular recognition. This field aims at discovering the suitable chemical compounds that would inhibit (or sometimes catalyze) a certain molecular interaction. The basic idea for molecular inhibition is to find a suitable chemical compound that would bind to the target under study, and either block that site of interaction, denying the original partner from binding to the target, or induce a structural change in the target, making it impossible for the binding partner to bind to the receptor.

Available chemical space is vast and screening it entirely is not feasible at the moment (or in the near future). It is estimated that the number of small organic

molecules constituting the chemical space is more than 10^{60} [87]. Out of this large space, the Chemical Abstracts Service[5] at the American Chemical Society[6] currently reports around 55 million registered organic and inorganic chemical molecules. Practically then the chemical space to be looked at is infinite since the number of synthesized and characterized compounds is extremely small in comparison to that space. Our inability to sample the chemical space (by synthesizing and testing every possible chemical compound) presents the major challenge for drug discovery. However, even with the already sampled 55 million compounds, screening these compounds for every target is a very expensive procedure and thus unattainable. These obstacles pave the way for the incorporation of informatics into drug discovery to help cover the sampled chemical space.

1.3.1 Compound Libraries

In drug discovery, the terms "lead-like" and "drug-like" are often used to describe chemical compounds. Small molecules that bind to a target are often called "leads". These small molecules are used as starting points for building larger molecules with a tighter interaction to the target. Drug-like molecules are those that have potential to be later used as pharmaceuticals. This potential is quantified by several properties, mainly absorption, distribution, metabolism, excretion, and Toxicity (ADMet) [88]. These properties affect the drug-potential of a molecule. Absorption is important so that the drug will function upon oral (or sometimes intravenous) administration. Distribution of the drug is necessary for the molecule to reach its designated binding

[5] http://www.cas.org/
[6] http://portal.acs.org/portal/acs/corg/content

site. Very fast metabolism of the drug leads to inactivity of the molecule. Excretion of that drug is important to avoid the accumulation of foreign substance in the human body. Finally, toxicity is an important factor in determining the drug potential of a molecule to avoid further complications or patient death.

Compound databases are crucial to any computational drug discovery program. In order to try to bind small ligands to a certain target, you need 3 dimensional structures of these compounds. Many databases are currently available to deliver this service. ZINC [89] houses over 13 million purchasable compounds in ready to dock 3D formats. CHEMCATS[7] reports over 41 million commercially available products from 1215 catalogs (11.75 M unique compounds). Our in-house compound database, EDULISS [90], houses over 5.5 million compounds (over 4M are unique). These compounds can be filtered by common drug-like parameters like the Lipinski rule of 5 [91], the Oprea filters [92, 93], and the Astex rule of 3 [94]. Options are included for similarity and pharmacophore searches.

A lot of research has been invested into the development of drug-like or lead-like profiles for chemical compounds. These profiles provide a means of fast prescreening of chemical compounds prior to *in silico* or *in vitro* screening in an effort to decrease the searchable space and confine it into a subspace that consists of useful compounds only. The most widely used profiles are the Lipinski rule of 5 [91], Oprea filters [92, 93], and Astex rule of 3 [94]. The Lipinski rule of 5 accepts

[7] http://www.cas.org/expertise/cascontent/chemcats.html

molecules with molecular weight <= 500 daltons, aLogP <= 5, hydrogen bond acceptors <= 10, and hydrogen bond donors <= 5. The Oprea filter accepts molecules with molecular weight <=460 daltons, aLogP between -4 and 4.2, hydrogen bond acceptors <= 9, hydrogen bond donors <= 5, rotatable bonds <= 10, and rings <= 4. Finally, the Astex rule of 3 is more selective, accepting molecules with molecular weight <= 300 daltons, aLogP <= 3, topological polar surface area <= 60 $Å^2$, and rotatable bonds <= 3. Of the 5.5 million compounds in the EDULISS database, 3.9 million compounds fit the Lipinski rule of 5, 3.4 million compounds fit the Oprea filters, and 520 thousand compounds fit the Astex rule of 3.

1.3.2 Protein Based Drug Discovery

As the title hints, protein based drug discovery aims at finding drug-like or lead-like molecules by studying the protein target itself. This method is particularly useful when there is a good knowledge base about the protein target and when the information about the experimental ligand is not very abundant. Two paradigms to perform this task are available: searching for binding molecules via sequence and structural alignment and searching for binding molecules via docking programs.

Similar proteins should have similar functions and thus, bind similar molecules [27]. This is the main drive behind sequence and structural alignment techniques leading to drug discovery. Sequence alignment methods (introduced in section 1.2.2) are capable of finding homologous proteins to the target. If the binding sites of these proteins are similar to that of the target, there is a high probability that molecules known to bind to these proteins would bind to target. This scenario is usually a good

option when the three dimensional structure of a protein is unknown. Another option is to use the sequence identity to predict (or model) the structure of that protein (with programs like MODELLER [95]). Then this theoretical structure is used for structural based approaches. Another similarity based option is using structural similarity to identify structurally homologous proteins (methods discussed in sections 1.2.3 and 1.2.4). These methods are capable of picking up structures with structural similarity to the target, regardless of sequence similarity. Molecules that bind to these homologous proteins are likely to bind to the target as well (especially if the binding sites are structurally similar). Molecules generated via similarity methods can also be used for ligand based drug discovery methods (section 1.3.3) as well.

Docking is the other option in protein based drug discovery. This option is only viable when a three dimensional structure of the target exists (or a high quality model). This option requires knowledge of the location of the binding site of the molecular interaction to be inhibited. It also requires a library of compounds to be used for virtual screening. Docking methods are discussed in section 1.2.5 and compound libraries are discussed in section 1.3.1. Chapter 2 of this work presents a novel algorithm (Surface Triplet Propensities, STP) that predicts the location of binding sites on proteins [96]. This helps in running docking simulations on targets where the interaction sites are unknown. Chapter 3 discusses the various applications of the novel STP method like generation of pseudo ligands for docking programs, prediction of the enzyme classes, and ranking protein-protein docking orientations. Chapter 4 discusses the spatial and chemical characteristics of binding sites and how

these qualities can be used in the detection of these sites and targeting them for drug discovery.

1.3.3 Ligand Based Drug Discovery

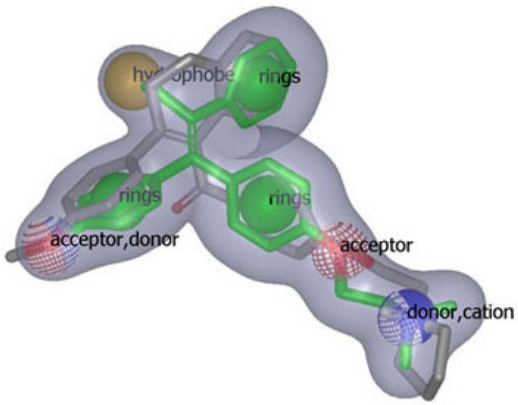

Figure 1-11: The mapping scheme used by ROCS in the fast molecular similarity searches. Figure retrieved from the ROCS official webpage[8].

Ligand based drug discovery methods rely on mimicking the ligand shape and interactions in order to select or create a list of possible inhibitors of a certain interaction. A few algorithms have been designed to perform shape similarity comparisons like ROCS [97] and UFSRAT [98]. ROCS uses the overall molecular volume to define the shape of a molecule in addition to atom type descriptors (Figure 1-11). UFSRAT also uses atom type descriptors, but defines the shape of molecule through a series of descriptors (mean, standard deviation, and skew) that define distributions of distances between atoms in a molecule (Figure 1-12). Using shape comparison methods, compound libraries can be screened for compounds that mimic an originally known binding molecule. Since similar molecules are expected to

[8] http://www.eyesopen.com/rocs

interact with the same target in a similar manner [97], these shape comparison algorithms constitute a powerful method to search for drugs and inhibitors of a certain interaction.

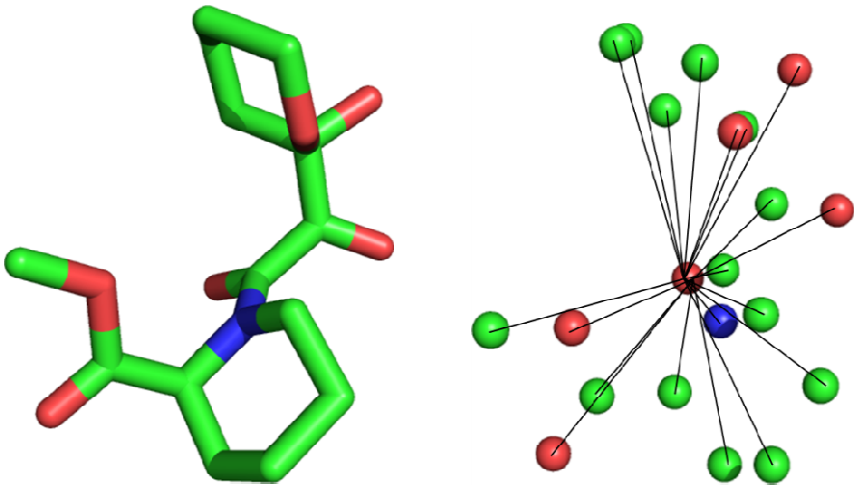

Figure 1-12: The UFSRAT method of defining distances between atoms to perform structural similarity calculations using the distribution of distances of all atoms to 4 key points: (1) the center of gravity of a molecule, (2) the closest atom location to (1), (3) the furthest atom location from (2), and (4), the furthest atom location from (3) [98].

Another ligand based method to inhibit a complex is to manually (or semi-automatically) design a list of compounds that mimic a certain interaction. These compounds are designed to have an overall shape, and refined to exhibit certain chemical features (donors, acceptors, hydrophobic, and acidic substructures) at special points, mimicking the same chemical features of the original ligand. Chapter 5 of this work describes a method that falls within this category of computational drug discovery approaches: designing compounds to mimic the β-turn motif of the Baff-Receptor ligand that binds to the B lymphocyte stimulator.

1.3.4 Interaction Databases

Another approach in computational drug discovery is building interaction databases. Such databases would house the known information about a certain type of interaction. Such an assembly of information provides a way to search for trends and motifs for that type of interaction. Examples of such databases include PepX [99] (protein-peptide interactions), GWIDD [100] (protein-protein interactions), and PDBbind [101, 102] (protein-small molecule interactions). PepX contains a nonredundant protein-peptide interaction dataset of 505 complexes. GWIDD (Genome-wide protein docking database) currently contains 25,559 experimental and modeled structures covering 771 organisms. Finally, PDBbind contains a total of 210 entries covering 70 different proteins, all annotated with experimentally measured binding affinity data.

ProPep [103] is an in-house protein-peptide interaction database. It consists of 481 non-redundant structures representing protein–peptide interactions where the protein is between 50 and 600 residues long and the peptide is between 3 and 50 residues in length. This dataset holds information about the protein and peptide residues involved in the interaction, the accessible surface area these residues (before and after binding), as well as the van der Waals and hydrogen bond interactions across the protein-peptide, protein-water, and peptide-water interfaces. Chapter 6 discusses the usage of this database to study the LxxL α helical motif. Using this database, we were able to pinpoint a structurally conserved LHRLL α helical motif that binds to nuclear receptors. Computational methods are then used to find inhibitors for this interaction.

1.4 Overview of this work

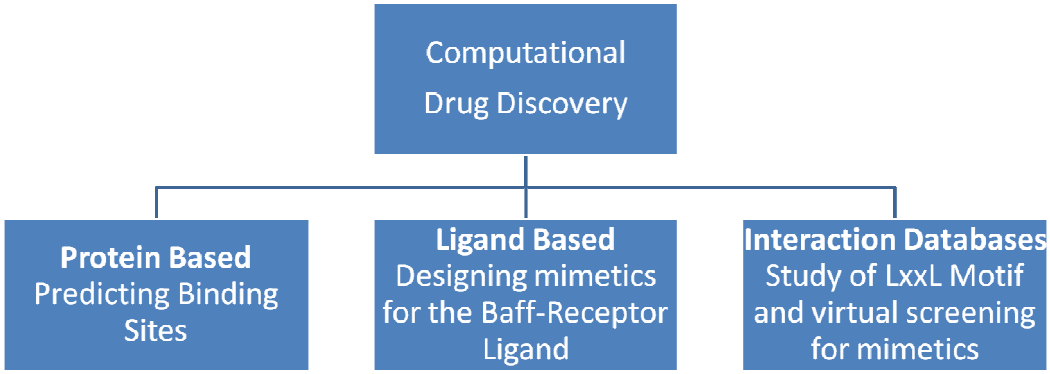

Figure 1-13: The contributions of this work to major sub-fields in computational drug discovery.

This work discusses various methodologies in computational drug discovery (Figure 1-13). Chapters 2, 3, and 4 focus on protein-based drug discovery, mainly with the prediction of the location of protein-ligand binding sites, the chemical and geometrical study of these binding sites, and the usage of information gathered about binding sites in various applications like predicting first level enzyme classification numbers. Chapter 5 focuses on ligand-based drug design, where the interaction between the B Lymphocytes Stimulater and the small ligand Baff-Receptor is studied, and cyclic hexapeptides are designed to mimic the structure of the ligand and hence inhibit that interaction. Finally, chapter 6 focuses on interaction database oriented drug design, where the protein-peptide database is used to study a peptide interaction motif (LxxL). This gave rise to computational efforts aimed at inhibiting the interaction of the LHRLL α helical motif with the nuclear receptors.

2 Surface Triplets Propensities

2.1 Introduction

2.1.1 Background

In silico experiments targeting protein interactions can be classified into several groups. Homology based algorithms use sequence or fold homology to infer protein function similarity. In 1991, Sander and Schneider [27] showed that a sequence identity of greater than 25% can infer structural similarity and hence, possibly functional similarity. Many studies [104-109] have successfully used sequence and structural similarity to characterize protein interaction. However, the fact remains that structure cannot always predict function (as in the case with TIM barrels that have the same fold but different functions [47]). Energy based algorithms focus on molecular mechanics and dynamics to try to find the most energy favorable conformation for a complex (protein-protein or protein-ligand). This type of algorithm requires a very accurate definition of the atomic forces (in type and value) and is very computationally expensive. Moreover, the difference between a good and a bad conformation is usually extremely small in magnitude, and hence this task becomes more difficult. Another type of algorithm is statistics based. These algorithms gather information on a certain type of interaction from all the structures that have already been crystallized in order to "train" the algorithms on predicting similar interactions. Such methods have utilized complex computational methods like Neural Networks [110-112] and Support Vector Machines [113-116].

At first, the in-silico algorithms faced a big problem: the lack of biological data that could be used to model molecular interactions. However, the structural databases have been growing fast, especially with the existence of structural genomics initiatives, aiming at filling the gaps in the databases by crystallizing "unrelated" proteins (usually less than 30% sequence similarity with currently crystallized proteins) [117]. With this expansion in the space of structural databases, especially by not focusing on a certain type or fold, in-silico algorithms can begin to explore the infinite universe of protein interactions.

An important application of studying protein-ligand interactions is its contribution to drug discovery. Finding compounds that can activate or deactivate certain proteins can mean finding a drug that would stop or slow down a certain disease. This process has two scenarios. The first is where the binding site of interest is known and the objective is to search for a ligand that would bind with relatively high affinity to this site (eg. FKBP12 [118], cyclophilin [119, 120]). The second is where the protein of interest is known, but the binding site on that protein is not and then, predicting the location of binding sites becomes crucial (eg. The complement immune system components 5 and 7 [121, 122] , thiosulfate sulfurtransferase (rhodanese) [123, 124]).

Several ways have been designed to search for protein binding surfaces. Some methods search for conserved regions and argue that conserved regions tend to be functionally significant [52, 125]. Other methods use force fields [51, 126] and

biochemical annotations and spatial motifs [127, 128]. A method has been designed to calculate the "Interface Propensities" of residues which indicates how frequently these residues tend to appear on binding sites [129]. This method was further explored to produce the PLB index (propensity for ligand binding) [130, 131]. However, calculating propensities for residues may yield noisy results since the packing and folding of proteins can bury some of the atoms in these residues beneath the protein surface. A residue based propensity might favor such buried atoms in appearing on binding surfaces. To avoid this noise, atom based propensity is calculated instead. This, however, introduces the problem of classifying protein atoms.

A classification of protein atoms into Atomic Groups was proposed by Tsai et al. [132]. Atoms are classified into 13 types called Atomic Groups based on heavy atoms, how many covalent bonds they have, and how many of these covalent bonds are actually with hydrogen atoms. Atomic Groups are denoted by XmHn, where X is a Carbon (C), Nitrogen (N), Oxygen (O), or Sulfur (S); m is the total number of covalent bonds X has, and n is the number of these bonds that are actually shared with a Hydrogen atom. These atomic groups have different radii and exhibit different biochemical properties. The interface propensities for these Atomic Groups is calculated and used to design an algorithm for predicting protein binding sites (Sections 2.3 and 2.4).

2.2 Surface Triplets Propensities (STP) algorithm

The STP algorithm is based on creating a score table for surface patterns, indicating the likelihood of a certain pattern to occur in a binding site. These score tables are calculated by training the algorithm on a certain representative dataset that represents an interaction type in the Protein Data Bank (PDB). Surface patterns are represented by triangles – triplets of atoms that can be simultaneously touched by a probe sphere of the size of a water molecule (Section 2.2.1).

Constructing such a score table from a dataset can be divided into several parts. First, a routine is needed to fetch the surface atoms and triangles from the proteins. A Second routine is needed to classify surface atoms between those that belong to the binding sites and those which do not. Finally, a third routine is needed to calculate the statistics and propensities of different atom/triplet types. Figure 2-1 summarizes this procedure.

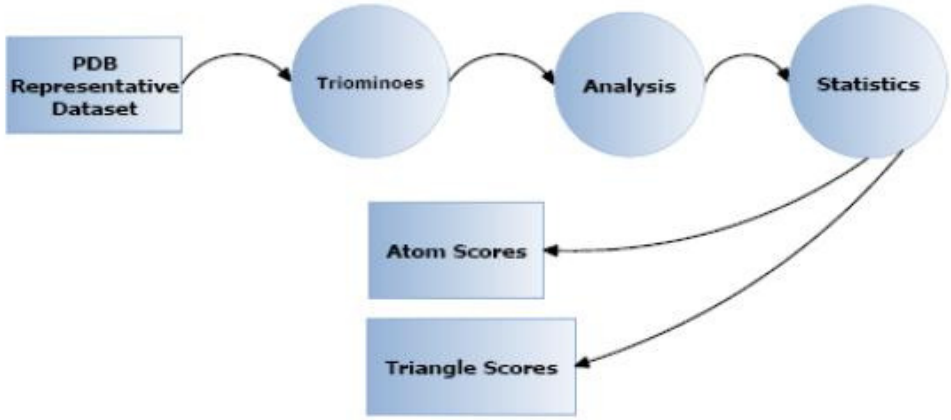

Figure 2-1: The framework of constructing the score tables from a representative dataset of the PDB starts by gathering surface triplets (Triominoes), finding out which triplets belong to a known binding site in that dataset (Analysis), and computing the propensities (Statistics).

2.2.1 Classification and triplet grouping of surface atoms

Protein atoms are classified into 13 atomic group types [132]. The atomic groups are defined according to the substitution pattern and number of covalent bonds on carbon, nitrogen, and sulfur atoms (Table 2-1). These atomic groups have different radii, electronegativity, and physiochemical properties. Consequently, searching for atomic group patterns on the surface of the protein would resemble searching for a physiochemical characteristic of that surface. Surface atoms are identified by rolling a water molecule (a probe sphere of radius 1.4Å) over the protein surface. A triplet (triangle) is defined as a group of 3 surface atoms that can be simultaneously touched by the rolling water molecule probe. Neglecting handedness, there are a total of 455 distinct 'triplet-types' that can be generated from combinations of the 13 different atom types. The "triominoes" program (written by Graham Kemp) takes in a PDB file and identifies the surface atoms and triplets. Figure 2-2 shows the triangles computed by the triominoes program. Triplets and atoms on the surface would be written out and used by other programs to continue the task of constructing the score table.

Table 2-1: The 13 atomic groups according to the classification of [132]. Atomic Groups are classified based on the heavy atoms (N, C, O, and S) and then subclassified by the number of covalent bond that heavy atom has and how many of these covalent bonds are with a hydrogen atom. Each of these atomic groups has a distinct radius and volume, contributing to unique physiochemical properties.

Atom Type	Example
N3H0	Pro N
N3H1	Amide N
N3H2	Arg NH1
N4H3	Lys NZ
O1H0	Carbonyl O
O2H1	Ser OG
C3H0	Carbonyl C
C3H1	Tyr CD1
C4H1	Ala CA
C4H2	Pro CB
C4H3	Ala CB
S2H0	Met SD
S2H1	Cys SG

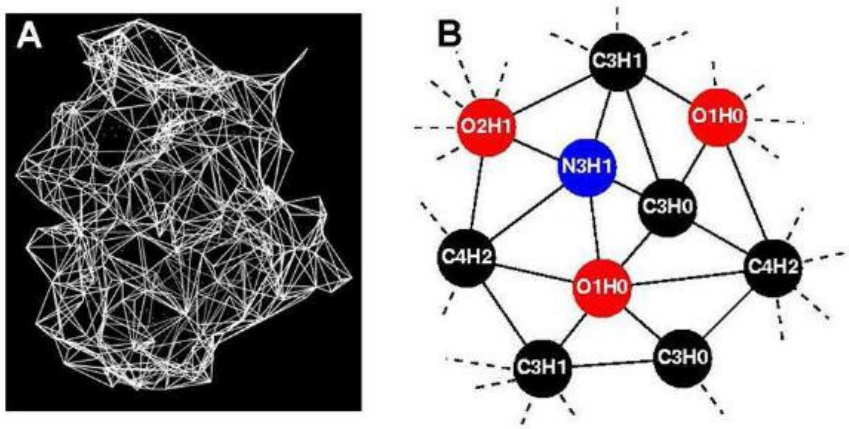

Figure 2-2: Triangles and Atoms on the Surface: A shows the general view of a protein structure, while B shows the details of the atomic classifications on such a surface. A triangle is identified as a triplet of atoms that is simultaneously touched by a probe sphere of the size of a water molecule.

A 14th atomic group was created for Arg amides (since it's not possible to distinguish the NH_2 (N3H2 atomic group) from the NH (N2H1 atomic group) of these residues in crystal structures). A subroutine to correct the types of sulfur atoms in disulfide bonds was also added. This subroutine would find all pairs of sulfur atoms which are less than 2.5 Å apart and change their type from S2H1 to S2H0. Atomic groups are ordered from 1 to 14.

2.2.2 Nomenclature

For computational reasons, the atomic groups are given special identifiers (Appendix Table 9-1). Triplets are named according to their constituting atomic groups. Since order is not important to our search, triplets made up of the same atoms in different permutations are considered to be the same. This classification gives rise to 455 different 'triplet-types'. A triplet is named according to its constituent atomic groups in an increasing order.

2.2.3 Classifying triplets and building score tables

The "analyze" program calls the triominoes program repeatedly to output two triangle files; the first containing triplets found on the entire surface of the protein and the second containing triplets found on the surface of the protein-ligand complex. Consequently, the first file will contain some atoms and triplets that are not found in the second file. These unique triplets and atoms have been concealed from the probe by the ligand upon binding, and therefore belong to the binding site between the protein and the ligand.

$$\text{InterProp}(\alpha) = \frac{\text{InterCount}(\alpha)}{\sum_{i=1}^{455} \text{InterCount}(i)} \qquad (1)$$

$$\text{SurfProp}(\alpha) = \frac{\text{SurfCount}(\alpha)}{\sum_{i=1}^{455} \text{SurfCount}(i)} \qquad (2)$$

$$\text{Calculated Propensity}\,(\alpha) = \log_2\left(\frac{\text{InterProp}\,(\alpha)}{\text{SurfProp}\,(\alpha)}\right) \qquad (3)$$

Where:
α designates one of the 455 possible triplet type;
InterProp(α) is the proportion of all ligand-binding-site triplets that are of type α;
SurfProp(α) is the proportion of all surface triplets that are of type α;
InterCount(α) is the count of occurrences of triplet type α in ligand binding interfaces in the dataset;
SurfCount(α) is the count of occurrences of triplet type α on protein surfaces in the dataset;
i spans the 455 triplet types.

Equation 2-1: Calculation of the triplet propensities. This calculation is only done once and the CalculatedPropensities are recorded in a score table. After that, the notion propensity refers to a value that is fetched from the score table.

The "statistics" program performs a simple statistical study to calculate the atom-propensities and the triplet-propensities that constitute the score table. The probability of finding a specific triplet in an interface and that of finding the same atom/triplet on the entire surface are calculated. The propensity of a triplet is

calculated as the logarithmic (base 2) of the ratio between these probabilities. Formally, the scores would be calculated according to Equation 2-1. For prediction, propensities are read from a scoretable that includes all the CalculatedPropensities for the 455 triplet types.

2.3 Constructing the STP Score Tables for the Protein-Ligand, Protein-Peptide, and Protein-Protein Interaction Datasets

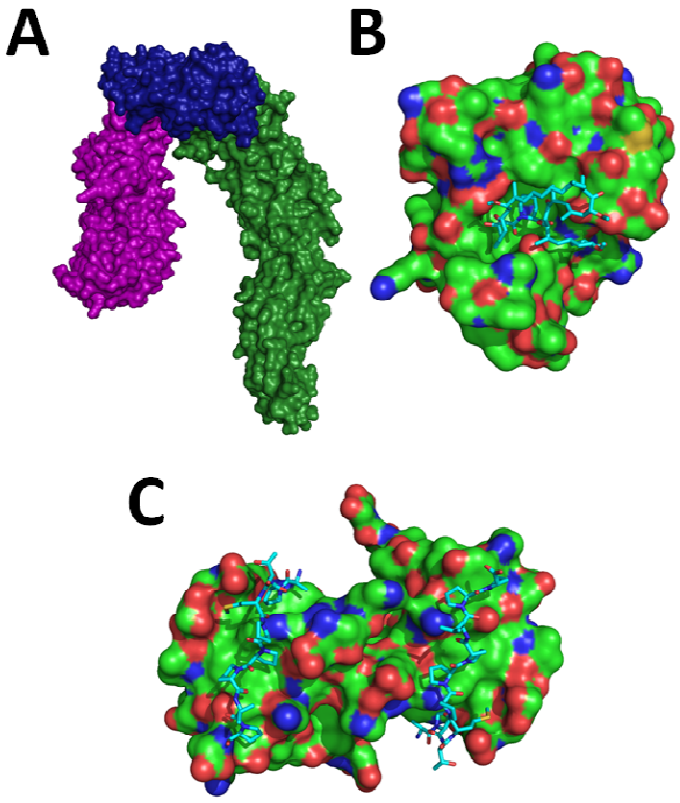

Figure 2-3: The classification of protein interactions into Protein-Protein (A, showing different proteins interacting to form the SCF E3 Ligase complex), Protein-Ligand (B, showing the binding of FKBP12 to the FK506 ligand molecule), and Protein-Peptide interactions (C, showing the ABL Tyr Kinase bound to 2 Pro rich decapeptides.

We refer to the interacting molecules as the receptor and the partner. Based on the type and size of the partner, protein interactions can be classified into three groups: protein-ligand, protein-peptide, and protein-protein (Figure 2-3). Peptides are defined as stretches of amino acids between 3 and 50 residues in length. Peptide sequences that are longer than 50 residues are classified as proteins. Ligands are defined as non-amino acid molecules having more than 10 carbon atoms which make at least 4 van der Waals interactions with the protein. Based on this classification, three score tables are created. The protein-ligand, protein-peptide, and protein-protein score tables are each constructed from an appropriate dataset (Sections 2.3.1 and 2.3.2). For any interaction type, the binding site is defined as the triangles on the surface of a receptor that have been concealed from a water probe upon binding to a partner.

2.3.1 The Protein-Ligand Score Table

A representative dataset of the Protein Data Bank (PDB) was compiled by Yi-Gong Sheng and stored in a database on the cycfs server under the name db_ecoli. This dataset was designed to represent all the known ligand binding sites found in the 2007 version of the PDB. This dataset is refined by using crystal structures with resolution better than 1.7 Å. A ligand is considered if it makes at least 4 van der Waals interactions with the protein, and if the maximum distance between any 2 atoms of this ligand is less than 23 Å. No two proteins in this test data set have a sequence identity greater than 50%.

The table JOB INFO in that database includes 352 protein-ligand interaction entries belonging to 316 different PDB structures and constituting a nonredundant

representative dataset of the protein-ligand interaction in the PDB. The coordinate files for these entries (in different formats like PDB, mol2, and sdf) are stored in the directory "/usr/cycfs/work/ecoli/lid_run/" and are divided into several folders. A file corresponding to an entry in the table JOB INFO is accessed through the value stored in the DIR column in the table JOB INFO. This value is the folder name inside "/usr/cycfs/work/ecoli/lid_run/" where the coordinate files are stored. Each directory includes 2 PDB files; one starting with the word "Prot" and this is a PDB file containing the coordinates of the protein and another file containing the coordinates of the ligand.

Table 2-2: List of PDB entries that make up the representative dataset of protein-ligand interactions (Yi-Gong Sheng's work). Some entries have multiple binding sites.

PDBID	PDBID	PDBID	PDBID	PDBID	PDBID	PDBID	PDBID	PDBID	PDBID
154L	1E4M	1GA2	1ID0	1KIC	1M7G	1O6G	1Q0R	1RGE	1UUY
16PK	1E6W	1GAI	1IE9	1KJQ	1M7Y	1O6I	1Q1A	1RRM	1UWC
1AJS	1E6Y	1GG6	1IN4	1KLL	1ME4	1O7G	1Q36	1RWH	1UXA
1AOE	1EEX	1GHE	1IS3	1KM6	1MFA	1O7J	1Q4U	1RXQ	1UY4
1AXW	1EJ0	1GKL	1IW0	1KMQ	1MG5	1O7Q	1Q74	1RYA	1UYY
1B0U	1ELU	1GM7	1IWH	1KMV	1MJH	1O8V	1Q92	1S1D	1V00
1B4P	1EN2	1GNX	1IYB	1KPF	1MP8	1O97	1Q9R	1S2A	1V0L
1B8O	1EU1	1GOR	1IYH	1KQF	1MR3	1OBD	1QD1	1SJW	1V2X
1BD0	1EVL	1GS5	1J1G	1KQR	1MRK	1OC2	1QGI	1SL4	1V3H
1BKF	1EWF	1GTV	1J1M	1KQW	1MV8	1OD6	1QH5	1SR7	1VHT
1BUP	1EXM	1GX5	1J1N	1KT6	1MWQ	1ODM	1QHO	1STY	1VHW
1BVD	1EYN	1H2B	1J54	1KWF	1MXG	1ODZ	1QJC	1SU2	1VK5
1BX4	1F0L	1H4E	1JA9	1L3L	1MXI	1OE8	1QJP	1T46	1VLB
1BXO	1F2U	1H61	1JAY	1L5O	1MZ9	1OFL	1QK3	1T7F	1W0P
1BYQ	1F5N	1H6H	1JBO	1L8N	1N08	1OGO	1QMG	1TAD	1W3L
1C1L	1F6B	1H8D	1JG1	1LC3	1N1T	1OH0	1QNR	1TBB	1WMS
1C4Q	1F74	1HFU	1JIF	1LJN	1N2E	1OI6	1QOP	1TBF	2MSB
1CCW	1F8E	1HMT	1JK3	1LKD	1N3Z	1OJJ	1QPC	1TH6	2NLR
1CG6	1F9V	1HNJ	1JKL	1LLF	1N5S	1OQ5	1QV0	1TX4	2TPS
1CRU	1FCY	1HP1	1JKX	1LO7	1N62	1OS6	1QW9	1U4B	3CHB
1CSH	1FK5	1HTW	1JP4	1LPC	1N6A	1OW4	1QXY	1U4G	3DFR
1CZQ	1FNC	1HX0	1JPZ	1LQT	1N83	1OWE	1QZ5	1U7G	3MAN
1D2S	1FP2	1HYV	1JTV	1LRI	1N8K	1P0H	1R2Q	1UDC	3MBP
1D3G	1FRB	1I0V	1JVP	1LUQ	1N8V	1P1J	1R5L	1UKV	3STD
1DAD	1FTK	1I12	1JX4	1LVW	1N9B	1P3D	1R5R	1UOG	4UAG
1DBW	1FZQ	1I1N	1JZ8	1LXK	1NB9	1P5Z	1R6D	1UR1	5P21
1DF7	1G0O	1I24	1JZI	1LZJ	1NF9	1PI5	1R6W	1URS	6CEL
1DIM	1G1T	1I3H	1K3Y	1M15	1NN5	1PIN	1R87	1URX	7ATJ
1DL2	1G2N	1I4F	1K4G	1M26	1NNF	1PJ6	1R8S	1US0	
1DZ4	1G3M	1I58	1KA1	1M2K	1NRJ	1PWB	1RA2	1UTP	
1E19	1G6H	1I76	1KB0	1M2R	1NSC	1PZG	1RDQ	1UU3	
1E2K	1G8K	1ICM	1KEI	1M4I	1NYW	1Q0N	1RFF	1UU6	

Most of the binding sites belong to unique PDB entries. Table 2-2 lists all the PDB ids as suggested by Yi-Gong's work. Most of the structures in this dataset included 1 ligand. However, some PDB entries include 2 or 3 binding sites, and these are shown in Table 2-3.

Table 2-3: PDB entries in the protein-ligand interaction dataset with multiple binding sites

PDBID	Binding Sites	PDBID	Binding Sites
1CRU	2	1MV8	2
1D3G	2	1N5S	2
1DL2	2	1N9B	2
1E4M	3	1NN5	2
1FCY	2	1O6I	2
1FP2	2	1OBD	2
1G3M	2	1OC2	2
1GA2	2	1OFL	2
1GTV	2	1OJJ	2
1HFU	2	1P3D	2
1HX0	2	1Q0R	2
1I58	2	1QH5	2
1JZ8	2	1QHO	2
1LKD	2	1QZ5	2
1M26	2	1R87	2
1M4I	2	1U4B	2
1M7G	3	1UYY	2

Due to the atom classification in the PDB structures, some entries were problematic to deal with. Many entries contained atoms with alternate locations. These cases were handled by writing a routine which takes in a PDB file, locates the atoms with alternate locations, and outputs the same file having adjusted it by keeping only those atoms with highest occupancy value.

For DNA-containing entries (1JX4, 1RFF, and 1U4B), a special routine was established to deal with these entries, omitting all DNA molecules, without misinterpreting the binding surface between the protein and the DNA as an outer surface. This allows for the characterization of the protein-ligand interface properly.

Some entries bind the ligands in internal pockets in the protein, totally inaccessible to the external environment of the protein. These cases represent binding sites cannot be characterized by the surface triplet approach and are therefore outside the scope of this study and have hence been deleted from the training set. These entries are: 1DZ4, 1QOP, 1KQF, 1L3L, 1R5L, and 1T7F. Moreover, entry 1GTV was deleted since the ligand in this structure was not characterized properly.

Finally, the asymmetric unit of entries 1P3D, 1JKX, and 1VHT contained duplicate structures (Figure 2-4). To avoid counting the same atoms twice, the "biological unit 1" for each of these structures was downloaded from www.pdb.org and used instead of the default model. With all these deletions and editing, the representative dataset of the protein-ligand interaction in the PDB was reduced to 309 structures.

The triplets and atomic groups on the protein surface and in binding sites were characterized as per Section 2.2.3. The propensities were calculated and are listed in Table 9-4 and Table 9-7.

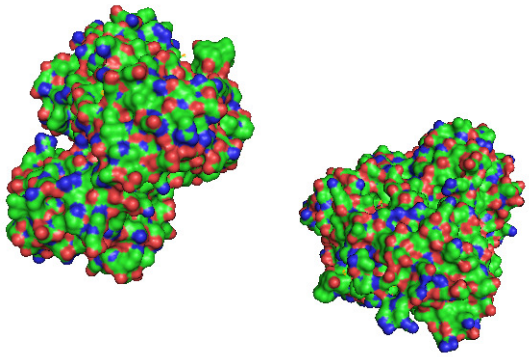

Figure 2-4: Crystal structure of the UDP-N-acetylmuramic acid (PDB id 1P3D), containing 2 duplicate structures. In such cases, the biological unit is used instead of the crystal structure.

2.3.2 The Protein-Peptide Score Table

Another score table was generated for Protein-Peptide Binding Sites. The Protein-Peptide Binding Sites score table was generated from the ProPep database generated by Simon D. Harding for his PhD studies [103]. The ProPep database holds a nonredundant representative dataset of the protein-peptide interactions in the PDB. The updated version of ProPep (updated by myself and reported in chapter 6) has been used and included 475 entries (Table 2-4). The score tables for this dataset are detailed in Table 9-3 and Table 9-6.

The dataset originally contained 481 structures. Five structures (1ATI, 1F59, 2A79, 2ZLD, and 3BFN) contained UNK amino acids in the protein and have been excluded. Structure 4Cpa included a GLX amino acid and has also been excluded. DNA entries were dealt with in a similar way like in the Protein-Ligand score table (Section 2.3.1).

Table 2-4: Protein-Peptide interaction dataset: PDB codes of the structures and the chains that contribute to the interfaces [103]. 148L:ES means pdb entry 148L, interface characterized as interaction surface between chains E and S.

PDBID	PDBID	PDBID	PDBID	PDBID	PDBID	PDBID	PDBID
148L:ES	1I3Z:AB	1O9U:AB	1T7R:AB	2A3I:AB	2G58:AB	2OX0:AC	2Z6W:BN
1A2X:AB	1I4F:AC	1OAI:AB	1TK2:AY	2A6D:AP	2G5L:AX	2OXW:AX	2Z8P:AB
1ABO:AC	1I73:AB	1OBX:AB	1TL9:AB	2A6I:BP	2G6Q:AB	2P0R:BE	2ZJD:AB
1AQC:AC	1IHJ:BC	1OFS:CD	1TP5:AB	2ADV:CB	2GHQ:BD	2P0W:BQ	2ZMI:AC
1AWQ:AB	1IID:AO	1OHE:AB	1TWQ:AP	2AI4:AB	2GKW:AB	2P1O:BC	2ZNE:AC
1B8D:BG	1ISQ:AB	1OJ5:AB	1TZS:AP	2AIJ:XP	2GPH:AB	2P1T:AB	2ZPY:AB
1B8H:BD	1J2J:AB	1OK7:BC	1U00:AP	2AK5:BD	2GPO:AC	2P54:AB	3APR:EI
1BC5:AT	1J2X:AB	1OM9:AP	1U7B:AB	2AQ9:AX	2GS6:AB	2PAV:PV	3B95:BP
1BH9:BA	1J34:AC	1ORH:AB	1UEF:AC	2ARR:AP	2H1C:AB	2PBK:BC	3BEJ:BF
1BR8:IP	1J4X:AD	1OU8:AC	1UGX:AB	2ASU:BA	2HAL:AI	2PC4:DH	3BEV:AC
1BX9:AB	1J71:AB	1OW6:CF	1UHE:AB	2AUC:CD	2HC4:AB	2PEH:AC	3BFQ:GF
1C9I:AC	1JD5:AB	1P16:BD	1UHL:AC	2AW6:AF	2HCJ:BA	2PHK:AB	3BIN:AB
1CKA:AB	1JDH:AB	1P22:AC	1UJ0:AB	2AWW:AC	2HFE:CD	2PIE:AF	3BL2:AC
1CLV:AI	1JET:AB	1P4N:AB	1UJK:AC	2AX3:AB	2HJL:AC	2PUY:BE	3BOO:AB
1CVR:AI	1JG3:AC	1P7W:AB	1UKH:AB	2AXI:AB	2HMH:AB	2PV1:AB	3BQD:AB
1CYN:AC	1JK4:AB	1PEG:AP	1UM2:AC	2AZM:BD	2HPL:AB	2PVF:AB	3BQO:AB
1CZY:AD	1JMT:AB	1PFB:AB	1UMW:AE	2B1N:AB	2HU2:AB	2Q3C:AB	3BRD:AD
1D4T:AB	1JOT:AB	1PPM:EI	1UPK:AB	2B2W:AD	2I04:BD	2Q3Y:AB	3BS4:AB
1DDV:AB	1JW6:AB	1PQ1:AB	1URL:AB	2B3G:AB	2I3H:BD	2Q5Y:CD	3BTS:AE
1DEV:AB	1JXP:AC	1PSA:AI	1UTI:AD	2B9H:AC	2I6O:AB	2QAC:AT	3BU3:AB
1DKZ:AB	1JYR:AL	1PVC:40	1V9T:BC	2BBA:AP	2IE3:CI	2QAS:AB	3BUX:BA
1DOW:AB	1K3A:AB	1PZL:AB	1VC3:BA	2BE1:BD	2IFR:AB	2QBW:AB	3C0T:AB
1E6I:AP	1K5N:AC	1Q3L:AP	1VDN:AB	2BEC:AB	2IHS:BD	2QBX:BD	3C3R:AB
1EBD:BC	1K8D:AP	1Q3P:AC	1VGK:AC	2BEW:AC	2IQ6:AB	2QFC:BD	3C5I:CE
1EE5:AB	1KAP:PI	1QC6:BD	1VPP:VX	2BEZ:CF	2ISQ:AB	2QIC:AB	3C5T:AB
1EEO:AB	1KHQ:AI	1QD6:CA	1VYT:AE	2BPA:13	2ITK:AB	2QIY:AC	3C66:BD
1EG4:AP	1KIL:AE	1QJJ:AB	1W70:AC	2BQZ:EF	2IUH:AB	2QKH:AB	3C9Q:AL
1ELR:AB	1KJM:AP	1QLS:AD	1WBP:AB	2BR9:AP	2IWB:AB	2QME:AI	3CB8:AB
1ELW:BD	1KJV:AP	1QNG:AD	1WBX:AC	2BRQ:AC	2J32:AB	2QN6:AC	3CBL:AB
1EMU:AB	1KRL:BA	1QO3:AP	1WKR:AI	2C23:AP	2J6F:AC	2QOS:CA	3CDW:AH
1EVH:AB	1KUG:AB	1QQD:AC	1X3Z:AI	2C3I:BA	2J7Y:AB	2QT5:BY	3CFS:BE
1EZX:AB	1KYF:AP	1QRP:EI	1XHM:AC	2C5K:TP	2J9A:AD	2QXV:AB	3CM8:AB
1F1J:AC	1L6X:AB	1QTX:AB	1XIU:BF	2C5W:BA	2JAC:AB	2QZO:AC	3CQW:AC
1F47:BA	1LB6:AB	1QXA:AB	1XOC:AB	2C74:BQ	2JAM:AD	2R0Y:AB	3CS0:AB
1FAV:AC	1LK2:AP	1QZ2:AG	1XRP:AQ	2CI9:BM	2JBY:AB	2R5M:AL	3CVP:AB
1FIV:AB	1LKK:AB	1R17:BD	1Y2A:CP	2CIA:AL	2JD5:BC	2REM:CT	3D24:CD
1FJM:AM	1LM8:VH	1R1Q:AC	1Y43:BA	2CNZ:AB	2JDL:BD	2RFI:AP	3D3X:BD
1FPR:AB	1LOP:AB	1RBD:AS	1Y7L:AP	2D7C:AC	2JGB:AB	2RHI:AB	3D7V:AB
1FQJ:AC	1LQV:AC	1RDQ:EI	1YC5:AB	2DOH:XC	2JK9:AB	2RI7:AP	3D8C:AB
1FYN:AB	1LVB:AC	1RFF:AC	1YCQ:AB	2DRK:AB	2JKG:AP	2RIV:AB	3D9T:BD
1FZM:AP	1M1D:AB	1RJL:CD	1YDP:AP	2DS2:DC	2NL9:AB	2RKY:AD	3DB3:AB
1G0Y:RI	1M45:AB	1RXZ:AB	1YMT:AB	2DVS:BQ	2NM1:AB	2RMP:AB	3DD7:CD
1G1S:AD	1MA3:AB	1RZX:AB	1YP1:AB	2EAX:CL	2NNU:AB	2UWJ:GF	3DEP:AB
1G6G:AE	1MF4:AB	1S4V:BD	1YTI:AI	2EGN:AB	2NPM:AX	2UXN:AE	3DIW:AC
1GA6:AI	1MFG:AB	1S5P:AB	1YUC:AC	2EQ7:BC	2NUD:BD	2V2F:FA	3DS4:AT
1GCT:AB	1MIZ:BA	1S6C:AB	1YWO:AP	2F31:AB	2O02:AP	2V3S:BC	3DSF:HP
1GEC:EI	1MTP:AB	1SCN:EI	1YWT:BD	2F4O:AI	2O40:AI	2V3Z:AB	3DXE:CD
1GFF:13	1MZW:AB	1SDX:AE	1YYE:BD	2F69:AB	2O5G:AB	2V6Q:AB	3DY0:AB
1GTJ:24	1N0W:AB	1SE0:AB	1YYP:AB	2F8E:XA	2O88:AC	2V8C:AC	3E2B:AC
1GUX:BE	1N12:CD	1SEM:AC	1Z8G:AL	2F9D:AP	2O9V:AB	2V8Y:AB	3E5A:AB
1GVU:AI	1N5Z:BQ	1SHA:AB	1ZBC:AC	2FF3:AC	2OBH:BD	2VGO:BC	3EGD:BD
1H0G:AD	1N7F:BD	1SOZ:AD	1ZGX:AB	2FF6:AH	2OD8:AB	2VIF:AP	3EMH:AB
1H27:BE	1NKM:AC	1SSH:AB	1ZGY:AB	2FFF:BA	2ODB:AB	2VKN:AC	3EMW:AB

1H6W:AB	1NLN:AB	1SUA:AC	1ZH7:AC	2FFU:AP	2OJU:AC	2VM6:AB	3MAT:AI
1H9O:AB	1NLT:AB	1SVF:AB	1ZJ7:AI	2FGR:AB	2OKR:AC	2VPB:AB	5TMN:EI
1HC9:BD	1NQ7:AB	1SVZ:AC	1ZOQ:BD	2FIB:AB	2OM2:CD	2VSL:AB	
1HSB:AC	1NRL:AC	1SZA:BZ	1ZVZ:AB	2FLU:XP	2OQ1:AB	2VZG:BA	
1Htr:BP	1NTV:AB	1T0J:BC	1ZY1:AD	2FMK:AB	2ORZ:AB	2Z23:AB	
1HUC:BA	1O6L:AC	1T15:AB	1ZYS:AB	2FOJ:AB	2OVH:AB	2Z3C:AI	
1I31:AP	1O9D:AP	1T6O:AB	2A25:AB	2FTS:AP	2OVR:BC	2Z3N:AC	

2.3.3 The Protein-Protein Score Table

A third score table, for Protein-Protein Binding Sites, was produced during my masters studies [133]. This score table was generated from a dataset created by Keskin et al. [134]. Table 2-5 identifies the binding partners listed in that dataset after all the necessary adjustments were made (removal of theoretical models and NMR ensembles, replacement of obsolete structures, and usage of biological models when the asymmetric unit includes 2 duplicate structures). The score tables for this dataset are detailed in Table 9-2 and Table 9-5.

This dataset was used to construct a protein-protein score table as part of my masters studies [133]. The dataset was designed by Keskin et al. (2004) [134] as a representative dataset of protein-protein interactions in the PDB, comprising 295 structures (Table 2-5). Structures 1BZI, 1CDA, 1DF2, 1I15, 1IF3, 1K9N, 1LT2, 1MLP, 1PAI, 1RLX, and 2BU0 are theoretical models and were omitted. Structures 1A7F, 1AZE, 1DT7, 1JEG, and 1JWD are NMR models and hence model 1 from each structure was used. The NMR models 1A0N and 1BON had corresponding mean models in structures 1AZG and 1BOM respectively, and those counterparts were used. Structure 1SEB was omitted for containing UNK residues. Structure

1G6R had 2 duplicate structures and was substituted with the biological model 1.

DNA containing entries were also dealt with as per Section 2.3.1.

Table 2-5: Protein-Protein interaction dataset: PDB codes (first 4 characters) of the structures and the chains (last 2 cahracters) that contribute to the protein interfaces [134]. 10GS:AB means pdb entry 10GS, interface characterized as interaction surface between chains A and B.

PDBID	PDBID	PDBID	PDBID	PDBID	PDBID	PDBID	PDBID
10GS:AB	1AZE:AB	1CSE:EI	1F3J:AP	1FNT:KZ	1HYR:AC	1JJO:CE	1PMA:BY
1A0N:AB	1B35:BC	1CU4:HP	1F3J:BP	1FNT:MN	1HZD:AB	1JK8:AC	1PMA:BZ
1A14:HL	1B48:AB	1CYD:AD	1F4M:AB	1FO0:AB	1I01:AB	1JK8:BC	1PMA:Z1
1A2P:BC	1B67:AB	1D0G:AB	1FAK:HI	1FO0:HB	1I10:AC	1JLV:AB	1PPF:EI
1A2Y:AB	1B77:AB	1D3B:AB	1FBY:AB	1FUU:AB	1I4K:12	1JS1:XY	1PSR:AB
1A6A:AC	1B9B:AB	1D3B:AB	1FI8:AC	1FZA:AB	1I4K:12	1JWD:AB	1QBZ:AC
1A6A:BC	1B9C:AB	1D5S:AB	1FJ1:BF	1G0U:OP	1I8F:AG	1JXZ:AB	1QD9:AB
1A6U:LH	1BEV:12	1D9K:AB	1FM6:DE	1G0U:OV	1IAK:AP	1K1F:DF	1QGH:AK
1A7F:AB	1BEV:13	1D9K:CP	1FNT:AB	1G1K:AB	1IAK:BP	1K4W:AB	1QMO:AB
1A8K:AC	1BEV:14	1D9K:DP	1FNT:Ad	1G3I:GM	1IC2:AB	1K6J:AB	1QU9:AB
1A8M:AB	1BEV:23	1DCI:AB	1FNT:AH	1G3I:GR	1IC2:CD	1KBA:AB	1RVF:14
1ABO:AC	1BH8:AB	1DEE:AD	1FNT:AI	1G6R:BH	1IES:BF	1KCG:AC	1RVV:12
1ABR:AB	1BJ1:HW	1DI0:AB	1FNT:Bc	1G7K:AB	1IJD:AC	1KD8:AB	1RYP:EL
1AC6:AB	1BJQ:AB	1DLH:AC	1FNT:BC	1G8Q:AB	1IO6:AB	1KEP:AB	1RYP:OP
1AG1:OT	1BON:AB	1DLH:BC	1FNT:BI	1G9I:EI	1IQA:AB	1KIL:AB	1RYP:RY
1AHW:EF	1BQP:AC	1DPS:AH	1FNT:BJ	1GCQ:AC	1IRJ:AB	1KIL:AC	1SBW:AI
1AIK:NC	1BSX:AX	1DT7:AB	1FNT:CD	1GK4:AB	1IRU:12	1KIL:BD	1SCJ:AB
1AKJ:DE	1BT6:AB	1DUB:AB	1FNT:Ci	1GL1:AI	1IRU:CK	1KJ4:AP	1SEB:EG
1AKM:AB	1BTM:AB	1DZ1:AB	1FNT:CJ	1GL2:AB	1IRU:EM	1KJF:AP	1SFC:BD
1AL2:12	1BX2:AC	1DZQ:AB	1FNT:CK	1GL2:AC	1IRU:FG	1KJH:AP	1SFC:BJ
1AL2:13	1BXK:AB	1E0B:AB	1FNT:DE	1GL2:BC	1IRU:FN	1KKQ:AE	1TAF:AB
1AL2:23	1BZX:EI	1E3S:AC	1FNT:Dh	1GNW:AB	1IRU:GH	1KQL:AB	1TAW:AB
1AO7:DE	1C08:BC	1E6J:HP	1FNT:DL	1GUY:AC	1IRU:I1	1KYO:BR	1TEC:EI
1AOH:AB	1C28:AC	1E8A:AB	1FNT:Eg	1GWC:BC	1IRU:J1	1L2I:AC	1TFX:AC
1AOI:AB	1C2Y:AB	1E92:AC	1FNT:EM	1H59:AB	1IRU:JK	1L7C:AC	1TGS:ZI
1AOI:CD	1C41:AB	1EAW:AB	1FNT:Ff	1HDC:AD	1IRU:KL	1LDT:TL	1TME:23
1AQD:AC	1C72:AB	1EBO:AB	1FNT:FG	1HEZ:AE	1IRU:KZ	1LGH:GJ	1TVD:AB
1AQD:BC	1C8O:AB	1EF7:AB	1FNT:FN	1HFO:AB	1IRU:OP	1LJR:AB	1YDV:AB
1AR8:14	1C9P:AB	1EJB:AE	1FNT:Ge	1HG3:AB	1IRX:AB	1LLD:AB	2AAI:AB
1AS4:AB	1CA7:AB	1EJM:AB	1FNT:GH	1HLE:AB	1JD1:AB	1LMK:AE	2MIP:AE
1AW1:AB	1CD0:AB	1EK6:AB	1FNT:HI	1HQK:AB	1JD2:LO	1MR8:AB	2SEB:BE
1AXC:AC	1CDT:AB	1EKX:AB	1FNT:HV	1HRI:12	1JD2:MN	1OTG:AB	2SIC:EI
1AXD:AB	1CE7:AB	1EQ2:AB	1FNT:Ia	1HRI:13	1JD2:NU	1PD2:12	2SIV:AB
1AYM:12	1COS:AC	1EZ4:AC	1FNT:Ja	1HRI:14	1JEG:AB	1PMA:12	2SNI:EI
1AYM:13	1COV:12	1F05:AB	1FNT:JK	1HVV:BD	1JFI:AB	1PMA:AB	3TMK:DG
1AYM:23	1COV:13	1F2D:BD	1FNT:JZ	1HWM:AB	1JH5:AB	1PMA:AC	6RLX:AB
1AZD:AC	1COV:14	1F2E:AB	1FNT:KL	1HYH:AB	1JI5:AC	1PMA:AP	

2.4 Predicting Binding Sites by Color coding the surface

The direct application of the collected atom and triplet propensities is to predict the location of binding sites. This function could be carried out by scoring patches on the surface of proteins. Patches that belong to the binding surface are expected to score higher than the other patches on the protein surface. Figure 2-5 outlines the procedure followed in order to predict the location of binding surfaces. The program that uses these propensities to predict the location of binding sites is called Surface Triplets Propensities (STP).

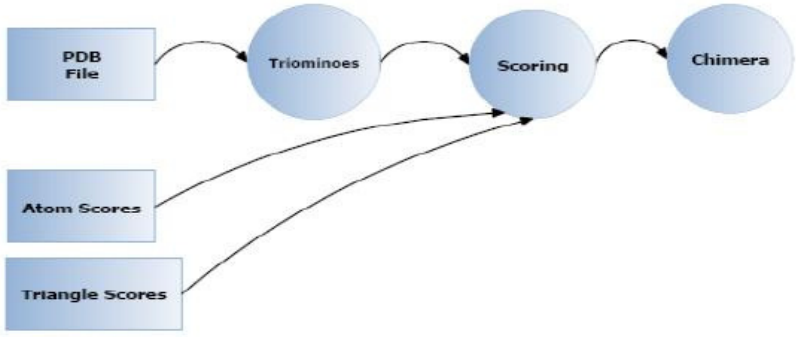

Figure 2-5: The binding site prediction process of STP. Triplets on the surface are extracted (Triominoes), then patches on the surface are scored using the score tables generated earlier (Scoring), and a visual output is created for different molecular viewing programs like PyMol, RasMol, and Chimera.

2.4.1 The 90-10 Testing Scheme and the Notion of Top Triangles

Computational Biology algorithms always face the challenge of proper and unbiased testing. The 90-10 testing scheme is used in all the testing routines. This scheme enables the testing of an algorithm with its own dataset, without any bias or redundancy. The training set of 309 proteins is divided into 10 mutually exclusive subsets, each constituting 10 percent of the set and comprising 31 structures. The

structures of each subset are tested with a score table generated from the remaining 90% of the original dataset. Since the dataset is nonredundant, this method of 10-fold cross validation is statistically correct and unbiased.

In some of the tests reported below, the notions of Top10, Top20, and Top30 are used. These correspond to the triplets with PatchScores greater or equal to 90, 80, and 70 respectively (when the PatchScores are scaled from 0 to 100).

2.4.2 Scoring Patches

The scoring process calls the triominoes program to create a list of atoms and triplets found on the surface of a specific protein. Each surface atom is then scored via a "sliding window" scheme. Each triplet is given a PatchScore which is defined as the average of the propensities of all the surface triplets whose centroids are found within a certain distance from this atom. This PatchScore would indicate the likelihood of an atom to belong to a binding site based on information from its surrounding atoms. The averaging function was used so that surfaces with more atoms on the surface are not favored over less-dense surfaces. A non-favorable patch might contain a highly favorable triplet and vice versa. The averaging function limits the effects of such noise on the general scoring. To better define a patch, Figure 2-6 illustrates a patch on a surface of a protein. Grouping the triplets into patches took the topology into consideration by creating a Graph of all the surface triplets, joining adjacent triplets with Graph Edges. This way, triplets that are spatially close but topologically distant (2 sides of a saddle structure) are not classified to be in the same patch.

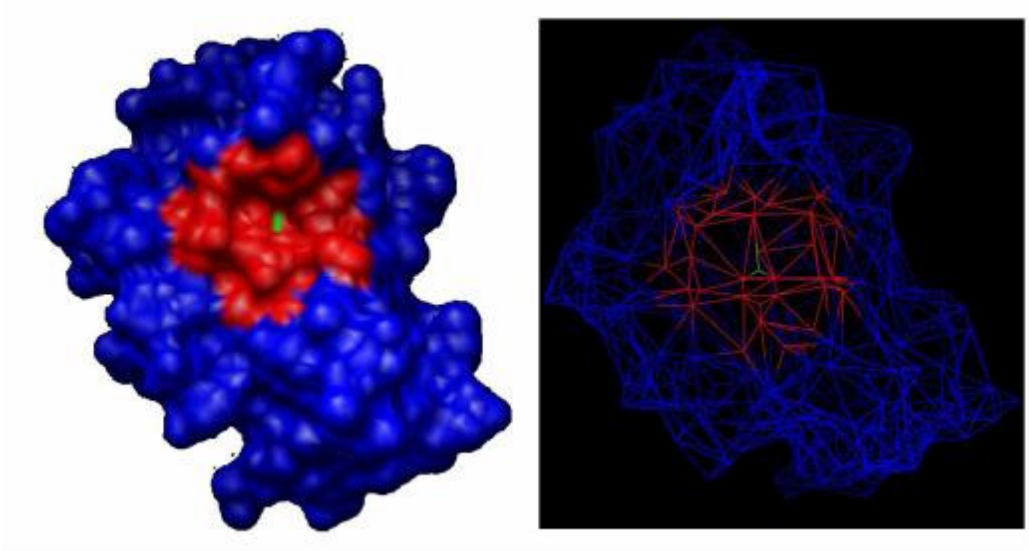

Figure 2-6: A Patch on the surface is shown in red surrounding the central atom highlighted in green. The picture on the left shows the solvent accessible surface while the one on the right shows the triangles computed by the triominoes program.

2.4.2.1 Color coding the surface

To give a visual output, PatchScores attributed to each atom are scaled from 0 to 100 (internal atoms are given a dummy PatchScore). Most molecular viewers are capable of using B-Factors (converted from PatchScore) to color the protein structure from blue to red (blue being the least favorable for a binding surface and red the most favorable). The coloring process then outputs scripts that can be used by visual programs like UCSF Chimera [44], RasMol [135], and PyMol [43] to color the surface of the molecule. It is informative to look at the coloring effect on the solvent accessible surface (solvent-excluded molecular surface as designed by [136]). This surface can be seen in Chimera and Pymol. In Pymol [43] for example, a user clicks "Show Surfaces" and then "color by spectrum, B-Factors" to obtain the colored surface.

2.4.2.2 The Effect of varying the Patch Diameter on Binding Site Prediction

The effect of varying the Patch Diameter was studied in order to determine the best value(s) for this diameter. Several tests were run to assess the output of the coloring program on diameter values ranging from 1Å to 40Å, at 1Å steps. In these tests, the notion of Top Triangles is used. A Top X triangle is a triangle with a scaled PatchScore (from 0 to 100) greater or equal to 100 – X. For example, a top 20 triangle should have a PatchScore of 80 or above on a 0 to 100 scale.

Three specific values were monitored: (1) the fraction of several Top Triangles categories within 5Å of a ligand (Figure 2-7), (2) the fraction of ligands that are within 5Å of at least 1 Top Triangle category (Figure 2-8) and (3) the difference in average propensity between the binding interface and the rest of the protein surface (Figure 2-9).

The "90-10" testing scheme was used (Section 2.4.1), dividing the dataset into 10 mutually exclusive testing sets of 31 structures. The remaining 278 structures were used to create the score table for the different triplet types. Each structure is then colored according to its respective score table created from the remaining 278 structures. We calculate a *fraction of triangles* value per structure, corresponding to the proportion of Top Triangles that are within 5 Å of any ligand atom. For example, if we have 25 Top 10 triangles (PatchScore 90-100) on structure X, and 20 of these triangles are within 5 Å of the ligand in structure X, then the *fraction of triangles* value is 20/25 = 0.8. The *average fraction of triangles* computed in Figure 2-7

correspond to the average value of all the *fraction of triangles* calculated for the 309 structures in the dataset. The same scheme was used to calculate the fraction of ligands next to a Top Triangle as well as the PatchScore difference.

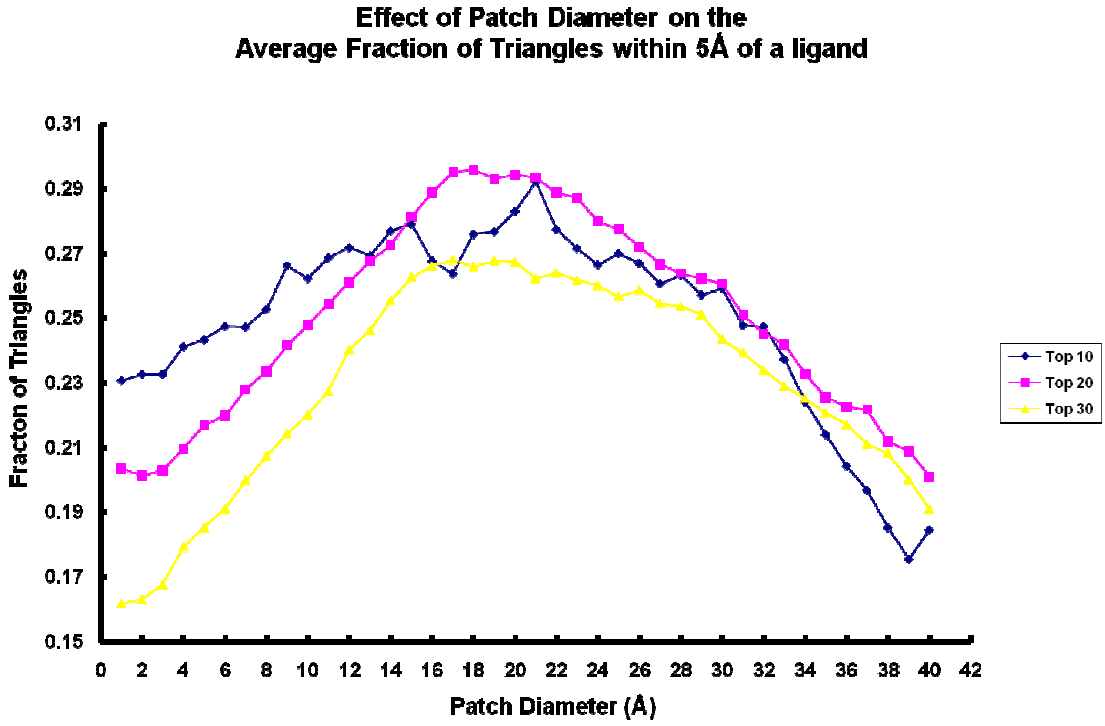

Figure 2-7: The Fraction of Top Triangles in the vicinity of a ligand (5Å from any ligand atom) as it varies with different coloring patch diameters. For each diameter, 309 *fraction of triangles values* are calculated (1 per structure), corresponding to the proportion of Top Triangles that are within 5 Å of any ligand atom. For example, if we have 25 Top 10 triangles (PatchScore 90-100) on structure X, and 20 of these triangles are within 5 Å of the ligand in structure X, then the *fraction of triangles* value is 20/25 = 0.8. The averages of these 309 values per diameter are plotted.

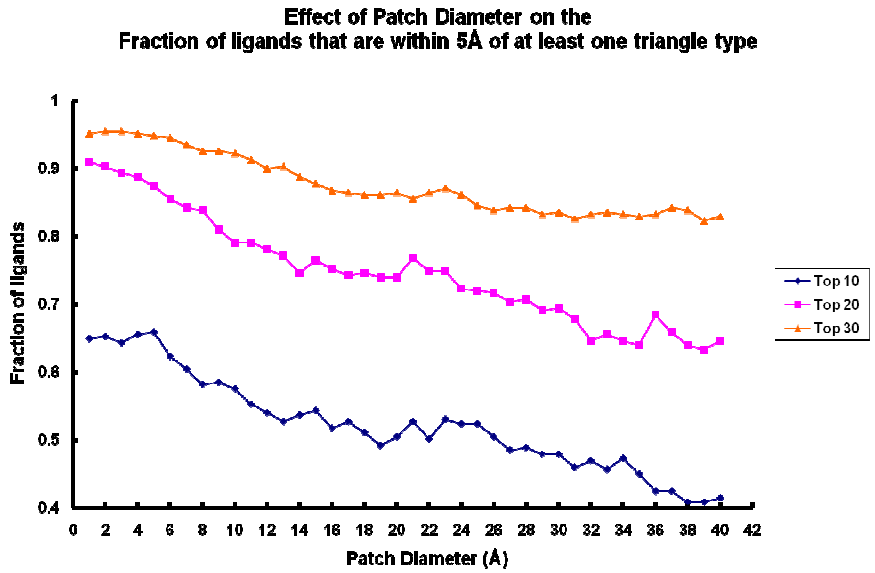

Figure 2-8: The Fraction of ligands in the Protein-Ligand interaction dataset that are in the vicinity of at least 1 Top Triangle (5Å from the triangle centroid). For each diameter, 309 true/false values were calculated, indicating whether the ligand is within 5Å from a Top Triangle. The plotted output is the fraction of structures having a "true" result

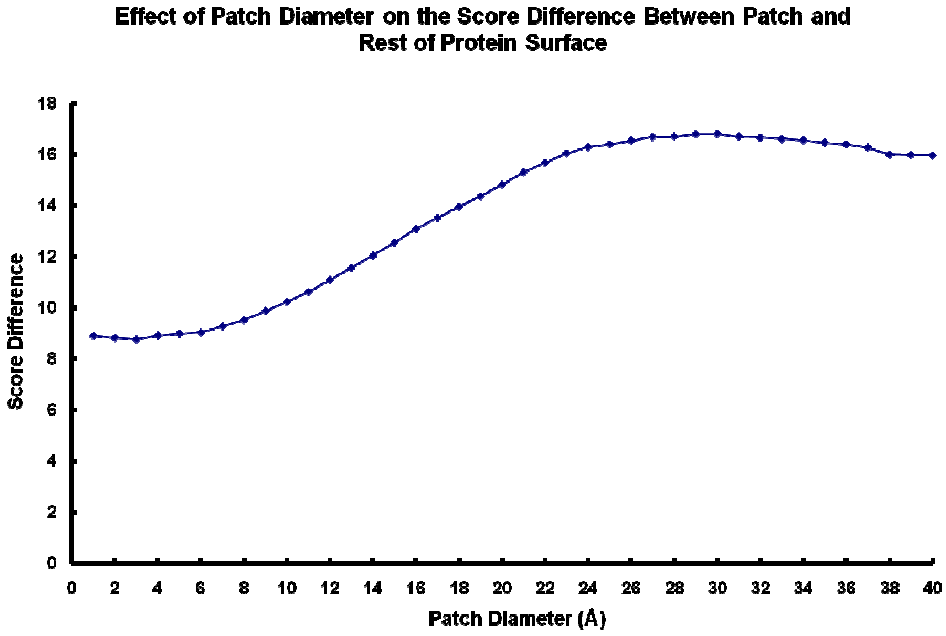

Figure 2-9: Distribution of the increase of average *PatchScore* between binding sites and Entire Surface as it varies with the patch diameter used to calculate the *PatchScores*. For each structure, PatchScores for all surface atoms are calculated (based on a certain patch diameter). Then the average *PatchScore* of all surface atoms is subtracted from the average *PatchScore* of all interface atoms. 309 increase values are calculated for every Patch Diameter and their averages are plotted.

As shown in Figure 2-7, there is a bell-shaped-like relationship between the patch diameter and the fraction of Top Triangles close to ligands, with the peak of the bell in the Patch Diameter range of 14Å to 26Å. That suggests that selecting the patch diameter in this range will provide the best possible coloring outcome. Figure 2-9 backs up this conclusion with the slope of the curve being at its highest in the 14Å to 26Å interval.

For patch diameters greater than 26Å, the difference in average propensities between the binding interface and the rest of the surface does not increase and its curve flattens. This means that increasing the patch diameter beyond the 26Å threshold does not enhance the binding site prediction quality. Figure 2-8 shows that as the patch diameter increases, the fraction of ligands in the vicinity of a high scoring patch decreases. That is expected since the lower the patch diameter is, the more the occurrence of small local hotspots. That gives ligands a higher probability to be next to one of those hotspots. It could be suggested that when searching for binding surfaces for smaller molecules, a smaller patch diameter could be used in the surface coloring. Larger molecules would essentially require a larger patch diameter. Figure 2-10 shows the changes in the surface coloring as the patch diameter varies.

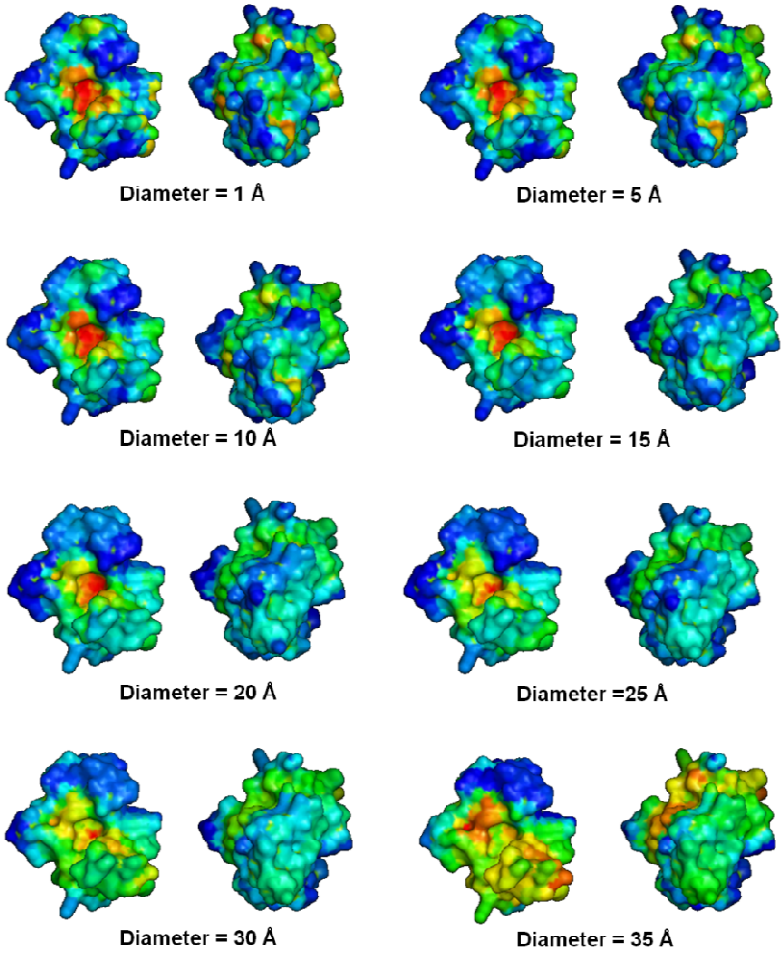

Figure 2-10: The effect of varying the patch diameter on the coloring output of STP. FKBP12 (PDB structure 2DG3) has been colored at different patch diameters. Each couple of pictures corresponds to the front and back faces of FKBP. The ligand binding site lies in the center of the front face (the left image in each group of 2)

The distribution of inter-triangular distances was also studied (Figure 2-11). Distances between any pair of binding site atoms (atoms that are concealed from the probe sphere upon the binding of a ligand) in a structure were measured. The Histogram in Figure 2-11 shows a normal distribution, with a peak in the range of [7Å – 11.5Å], giving more confidence to the choice of 15Å patch diameter (2 × 7.5Å).

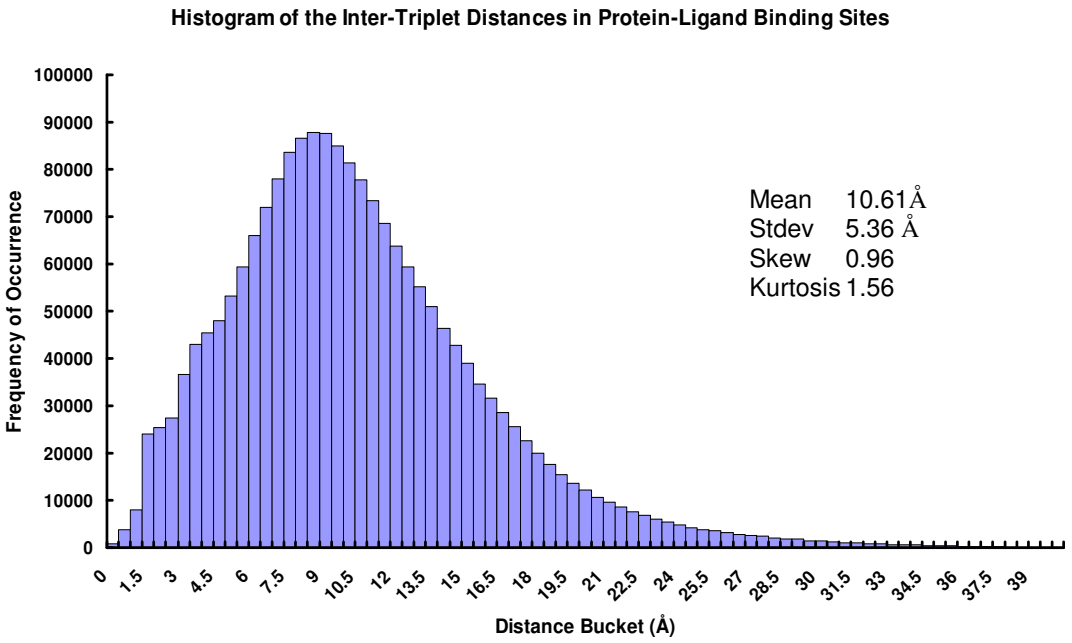

Figure 2-11: The Distribution of Inter-triangular distances within ligand binding sites according to the STP dataset shows a positively skewed bell shaped curve with a peak in the interval of [7.5 Å, 12 Å].

2.5 Testing the Ligand Score Table

2.5.1 Individual Cases

The algorithm has been tried out on several structures to assess its practical performance, ease of use, and accuracy. Those structures are introduced below and the STP predictions are shown.

2.5.1.1 FKBP12

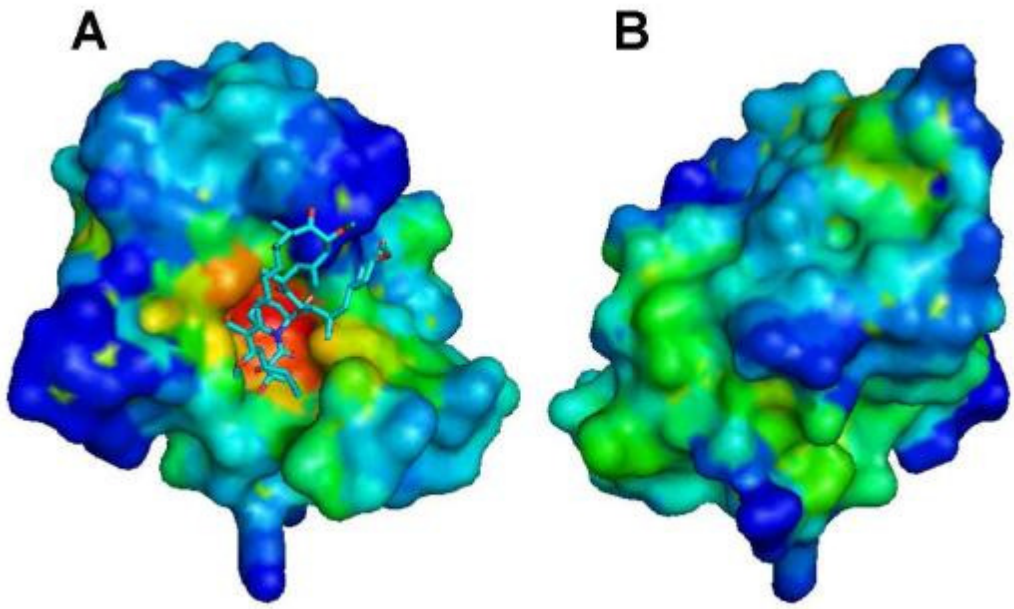

Figure 2-12: FKBP12 after being colored by STP. “A” shows the binding site and the ligand FK506, and “B” shows the back side of the protein.

FKBP12 belongs to a class of proteins known as the Peptidyl-prolyl cis/trans isomerases (PPIases) also known as rotamases and is involved in various molecular pathways like protein folding, signaling, trafficking, and transcription. Although the main function of FKBP class proteins is the activation of T-Cells (through the binding to FK506), this binding has side effects and implications in resisting various neurodegenerative diseases (e.g. Parkinson’s) [118]. Upon coloring with STP, the binding site of FKBP12 was predicted correctly (Figure 2-12). The PDB structure 2DG3 was used and colored according to the protein-ligand score table. The entire surface was colored from blue to green except for the binding pocket which was colored yellow, orange, and red. The program located the binding pocket successfully and without ambiguity.

2.5.1.2 The Gonadotropin-Releasing Hormone Receptor (GnRH-R)

GnRH-R belongs to the *Rhodopsin* G-Protein Coupled Receptor (GPCR) family. Its original ligand is a decapeptide which acts as a regulator of Luteinising Hormone (LH) and Follicle Stimulating Hormone (FSH); making GnRH-R an attractive target for controlling sexual functions [137, 138]. GnRH-R expression is also found in a number of carcinomas including breast and prostate [139, 140].

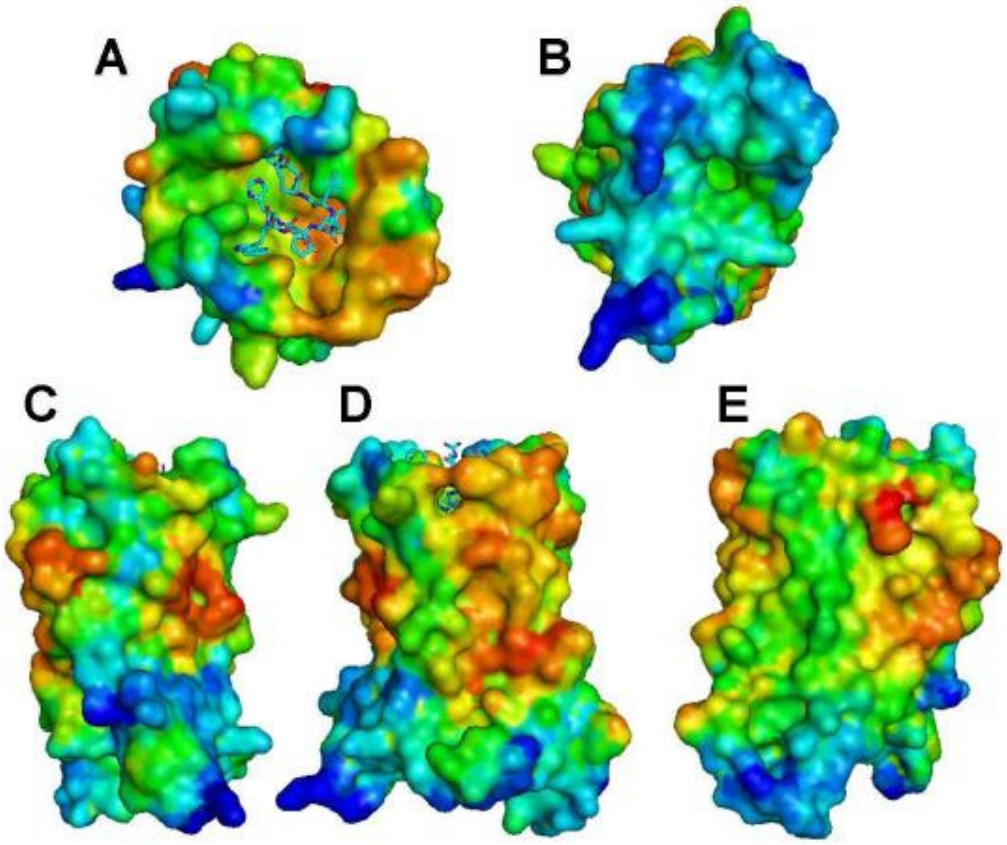

Figure 2-13: HGnRH after being colored by STP. "A" shows the top of the protein (extracellular part) with the bound ligand, "B" shows the bottom of the protein (intracellular part), and "C", "D" and "E" show the sides of the protein that correspond to the trans-membrane domain.

There is no crystal structure for this protein. It has been modeled based on the β_2 Adrenergic receptor (PDB ID 2R4R, 17% sequence identity, RMSD 2.751 Å) and

Rhodopsin (PDB ID 1U19, 16% sequence identity, RMSD 2.211 Å). An STP analysis was performed on this model and the results are shown in Figure 2-13. A large part of the GnRH-R surface is very hydrophobic since it is a transmembrane protein, and therefore the existence of the red patches on the transmembrane regions (Figure 2-13 C, D, E) is apparent. Two large cavities exist on the top and the bottom of the structure (Figure 2-13 A, B respectively), but only one of them was highly colored, and that was the one that was actually a ligand binding site. It is noteworthy that the colored half of the cavity is distorted in the inactive form of the receptor.

2.5.1.3 The ABL Tyr Kinase SH3 Domain

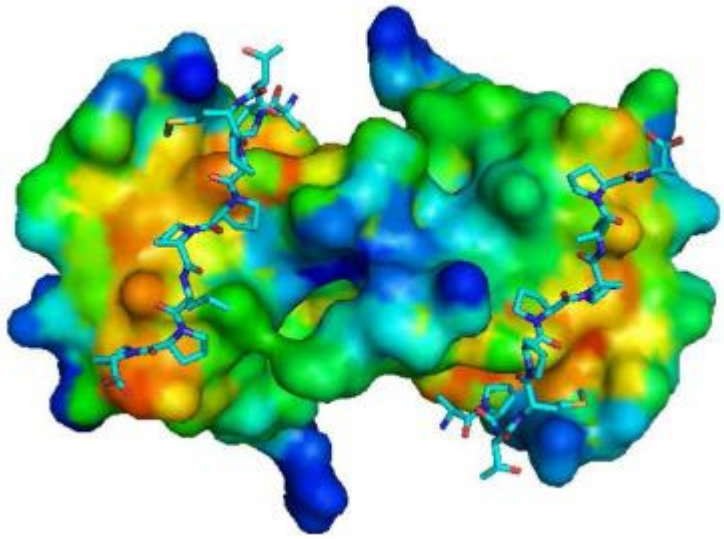

Figure 2-14: Abelson Leukemia virus tyrosine (ABL Tyr) Kinase SH3 Domain after being colored by STP showing the 2 shallow binding sites marked clearly.

The Abl-SH3 domain is linked to the down regulation of the Abl Kinase; binding to Pro-rich peptides having the PXXP motif or a PolyPro Type II helix conformation (PPII). This regulation takes place by interacting with the upper lobe of the Tyr Kinase, stabilizing it in the inactive form [141]. The PDB Structure 1ABO was tested

with STP and the two binding sites on the surface of the protein were unambiguously predicted (Figure 2-14). The strength of this prediction is that this binding site is shallow and would not be picked up by programs that search for geometric clefts on protein surfaces.

2.5.1.4 The SCF^{Cdc4} Ubiquitin Ligase

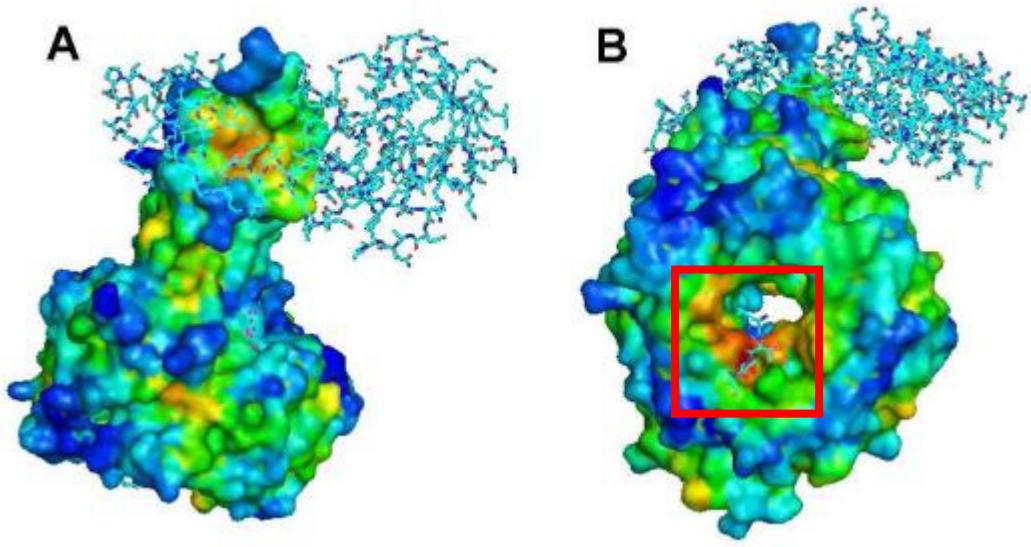

Figure 2-15: SCF^{Cdc4} Ubiquitin Ligase (PDB ID 1NEX) after being colored by STP. Two binding sites are discovered and highlighted by STP, the first (A) binds to the CBF3 subunit of the SCF^{Cdc4} Ubiquitin Ligase complex, and the second site (B) is the Cdc4-phosphodegron (CPD) Phosphopeptide recognition site.

The SCF^{Cdc4} Ubiquitin Ligase is responsible for the ubiquitination of the Cyclin-Dependent Kinase inhibitor Sic1, which in stable form is responsible for a G1 phase arrest [142]. The WD40 domain of the SCF^{Cdc4} (PDB ID 1NEX) has been colored with STP (Figure 2-15). STP predicted 2 distinct binding sites, one at the far side of the domain (Figure 2-15 A), and the other on the edge of a cavity at the other end (Figure 2-15 B). The first predicted site surface binds the CBF3 subunit of the

SCF^{Cdc4} Ubiquitin Ligase, while the second predicted site is the Cdc4-phosphodegron (CPD) binding site responsible for Phosphopeptide recognition, and is the most conserved part of the WD40 domain of the SCF^{Cdc4} Ubiquitin Ligase [142].

2.5.1.5 Human 11-β Hydroxysteroid Dehydrogenase Type 1

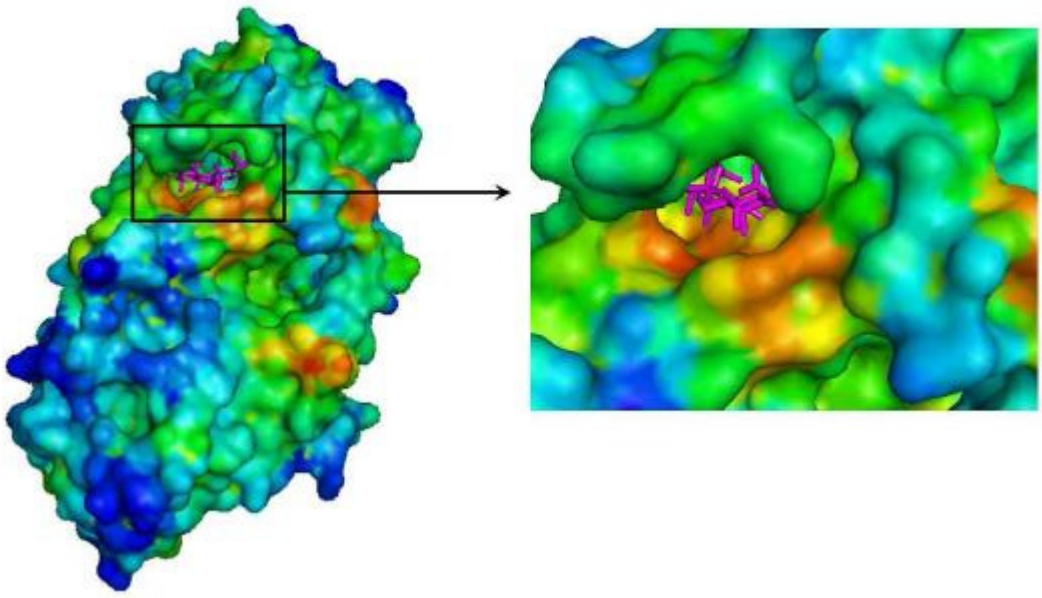

Figure 2-16: The Human 11-β Hydroxysteroid Dehydrogenase after being colored by STP.

The 11-β Hydroxysteroid Dehydrogenases (11β-HSDs) are responsible for converting glucocorticoids (like cortisone and dehydrocorticosterone) back and forth between their active and inactive forms. Unlike 11β-HSD type 2, the 11β-HSD type 1 (studied here) can catalyze this conversion in both directions. This enzyme plays a role in the hypothalamus-pituitary-adrenal axis, metabolic syndrome, and inflammation [143]. This enzyme was studied in the Walkinshaw lab by Jillian Adie

and Iain McNae. A crystal structure was produced and studied with STP (Figure 2-16), and the active site was successfully predicted. Two other red patches are visible on the surface, and they belong to the dimerization site of the protein.

2.5.2 Collective Testing and Validation

2.5.2.1 Propensity distributions

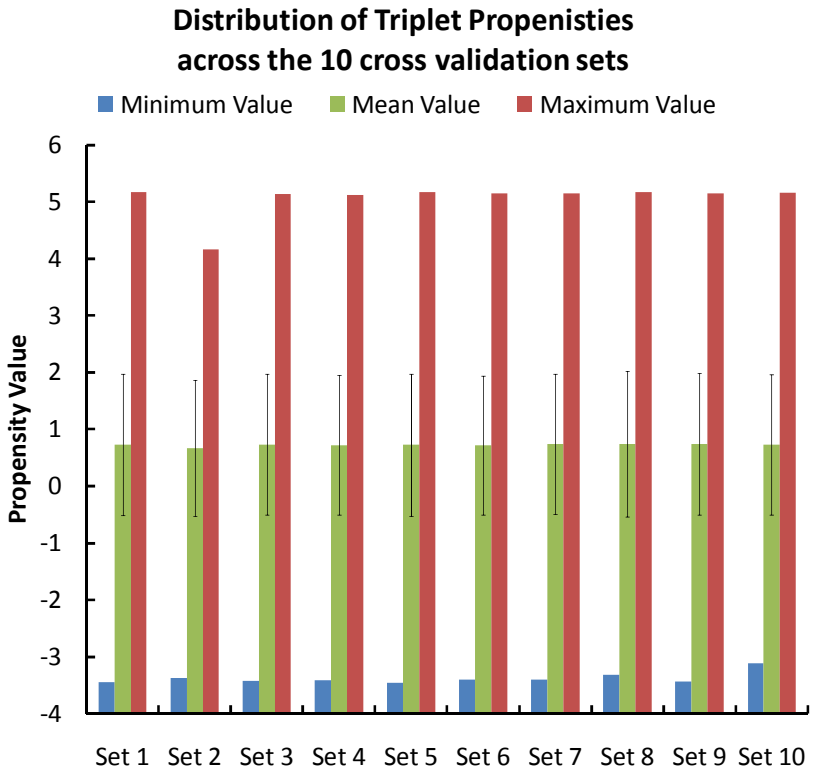

Figure 2-17: Distribution of the triplet propensities across the ten 90-10 testing subsets with the minimum (blue), maximum (red), average (green), and standard deviation (error bars) of the propensities of all triangular types. All ten subsets show similar attributes, confirming the consistent quality of STP scoring.

We studied the distribution of propensities in our dataset. The 90-10 testing scheme (Section 2.4.1) was used to avoid bias. The surface triplets on each structure in the

database were extracted and their propensities recorded. As per (Figure 2-17 and Figure 2-18), the distribution of each 90-10 subset was plotted separately. The distributions did not vary among the subsets and this is an indicator of the stability of the algorithm that maintained a constant performance across the 10 subsets. Approximately, propensities are in an interval between -3.5 and 5 with a mean of around 0.8 and standard deviation around 1.2. The trends shown across all 10 subsets indicate that the distribution is normal. That leads to 90% of the triplets occurring at a propensity between -1.77 and 3.22 (Mean ± [2 × Standard Deviation]).

Table 2-6: The 5 most frequent triplets and 5 least frequent triplets in the training dataset protein-ligand binding sites and their propensity scores. The occurrence corresponds to the number of times a triplet appears in the test dataset of 309 structures. The negative propensity of the most frequent triplets is not surprising as it means higher probability of occurrence, which decreases the propensity. The inverse is also true for rare triplets, where occurring once in a binding site will increase the propensity tremendously. Triplets that are abundant in binding sites are discussed in Table 2-7.

Triplet Type	Occurrence in Binding Sites	Occurrence on Protein Surfaces	Propensity
C3H0, C4H2, O1H0	796	65791	-1.21
C4H1, C4H2, O1H0	666	62708	-1.4
C4H2, C4H2, O1H0	663	54617	-1.2
C4H2, N3H1, O1H0	927	46342	-0.48
C3H0, C4H1, O1H0	452	36134	-1.16
C3H0, C3H1, S2H1	3	12	3.16
C4H1, S2H0, S2H0	1	14	1.35
O2H1, O2H1, O2H1	1	14	1.35
C4H1, O2H1, S2H1	2	15	2.25
C3H1, N3H2, S2H1	2	17	2.07

Table 2-7: The 10 most frequent triplets in binding sites in the training dataset protein-ligand binding sites and their propensity scores. The occurrence corresponds to the number of times a triplet appears in the test dataset of 309 structures.

Triplet Type	Occurrence in Binding Sites	Propensity
C3H1, C3H1, C3H1	951	2.24
C4H2, N3H1, O1H0	927	-0.48
C3H0, C4H2, O1H0	796	-1.21
C4H2, C4H3, O1H0	706	-0.36
C4H2, C4H3, C4H3	695	0.98
C4H1, C4H2, O1H0	666	-1.4
C4H2, C4H2, O1H0	663	-1.2
C3H1, C3H1, C4H3	638	2.09
C4H3, C4H3, C4H3	604	1.93
C4H1, C4H2, N3H1	598	-0.49

Only 10 triplet types lie outside this range in the score table generated from the entire dataset. Most of these triplets occur less than 200 times throughout the entire dataset (the average occurrence of a scoring triplet is 4000), except for two triplet types: (C3H0, N4H3, O1H0) (1493 occurrences and propensity of -1.86) and (C4H2, C4H2, N3H0) (998 occurrences and propensity of -2.29). The occurrence of these triplets suggests that although they are rare, they can still be found on protein surfaces (5 and 3 occurrences per protein structure respectively). Out of 1493 occurrences, only 12 of the (C3H0, N4H3, O1H0) triplet type were in a ligand binding site. Similarly, the second type exists 6 times in a binding site out of 998. Such triplets can be used as strong indicators to the nonexistence of a binding site since they are not extremely rare (25% of an average atom type occurrence) and exhibit a binding site propensity rate of only 0.9% and 0.6%. Listings of rare and highly abundant triplets are presented (Table 2-6 and Table 2-7).

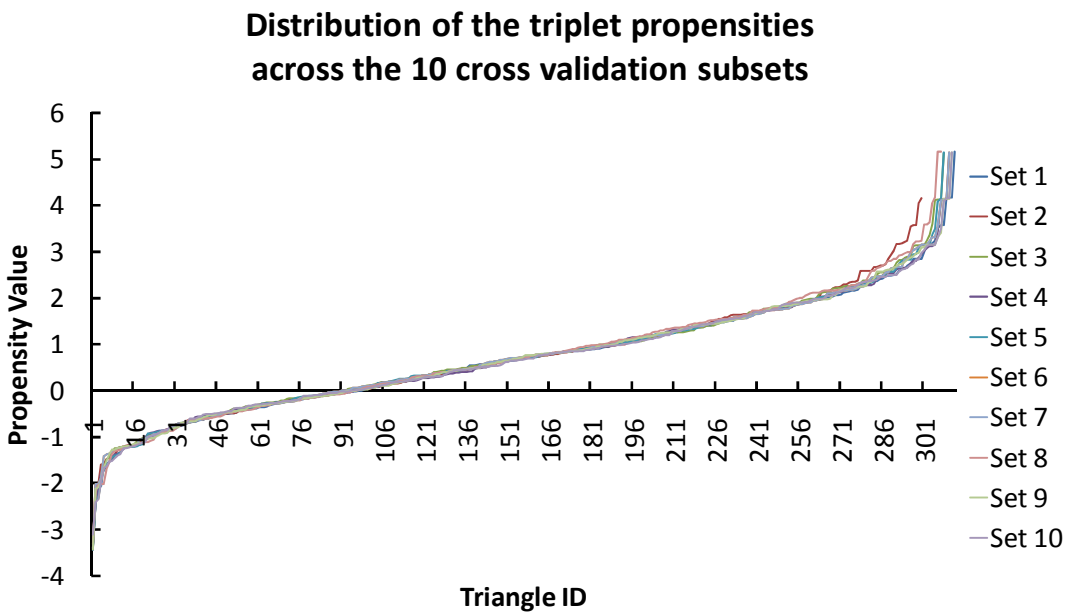

Figure 2-18: The distribution of triplet propensities across the ten 90-10 subsets. No major difference is noted.

2.5.2.2 Distinction of Binding Sites from their Surrounding Surfaces

This test aims to assess whether the STP score table yields a clear distinction between binding site triplets and entire surface triplets. A binding site is defined as a set of water accessible triplets that are concealed from the water probe when the ligand binds to the protein. For each set of the 10 fold cross validation sets 2.4.1), two distributions were recorded: distribution A containing the propensities of all surface triplets in a 90-10 subset, and distribution B marking the propensities of binding site triplets in that subset. A distribution shift was calculated as shown in Equation 2-2.

$$Shift(A,B) = \frac{Mean(B) - Mean(A)}{Stdev(A)}$$

Equation 2-2: Calculation of the shift between 2 distributions A and B.

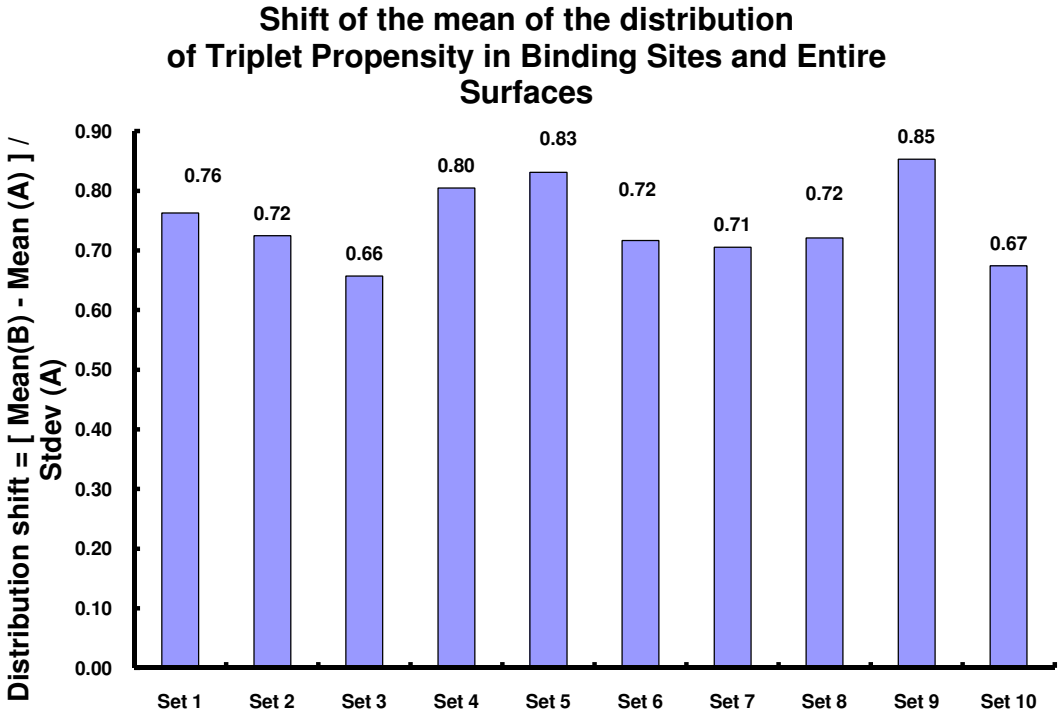

Figure 2-19: Distribution Shifts of triplet propensities across the ten 90-10 subsets of the Protein-Ligand interaction dataset. For each subset, the distribution shift is between the distribution of all binding site triplets propensities and the distribution of all surface triplets propensities in a certain structure.

All 10 subtests showed a consistency in the success of the algorithm. The distribution of propensities in the ligand binding sites had a higher mean than the entire surface distribution in general. The shift, when measured against the standard deviation of the entire surface distribution, ranged between 0.65 and 0.85 (Figure 2-19) indicating that triplets in binding surfaces are favored among the entire triplet distribution. This provides evidence for the success and power of the coloring scheme of STP.

Instead of testing the shift between individual triplet propensities in binding site and entire surface distributions, a new test was carried out to calculate such a shift between the distribution of average propensities. For each structure in a 90-10

structure, two attributes were recorded; the average propensity of binding site triplets and the average propensity of all surface triplets. The two distributions resulting from these two attributes were then tested and the distribution shift (Equation 2-2) was quantified (Figure 2-20 and Table 2-8). The average propensity of all triplets in the ligand-binding sites is higher (mean 0.32, standard deviation 0.41) and can be distinguished from average propensities for the entire surface (mean -0.3, standard deviation 0.08). The distribution shifts for the 10 subsets are in the range of 5.8 to 13 fold.

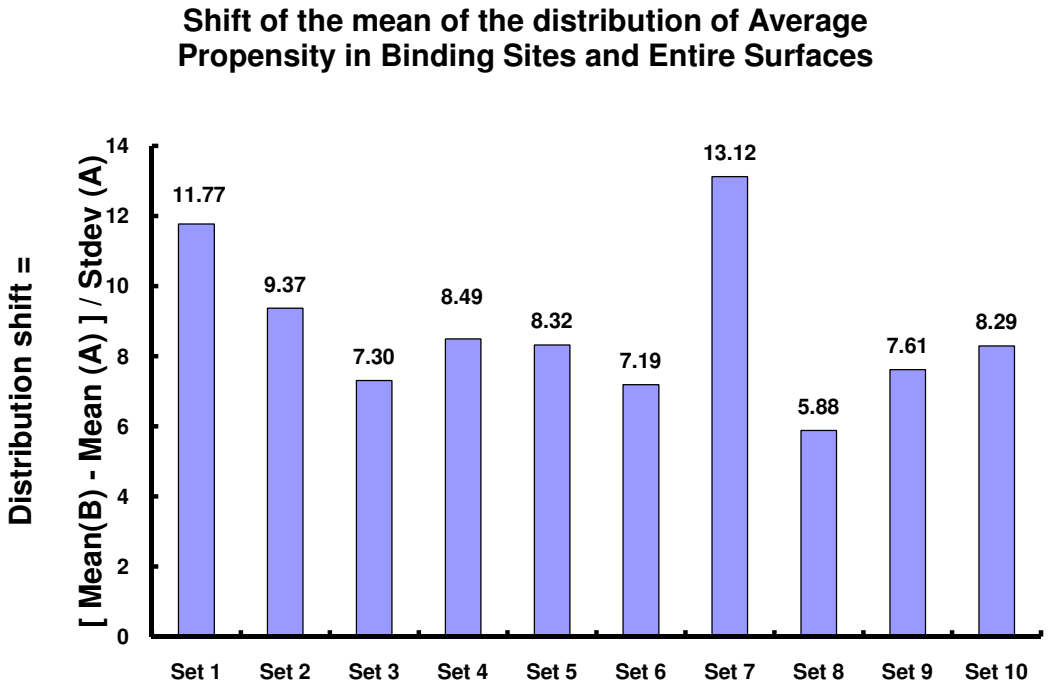

Figure 2-20: Distribution Shifts of propensities across the ten 90-10 subsets of the Protein-Ligand interaction dataset. For each subset, the distribution shift is between the distribution of the average of all binding site triplets propensities and the distribution of the average of all surface triplets propensities in a certain structure.

Table 2-8: Distribution of the average propensities of triplets in binding sites has a higher mean than the distribution of the average propensities of all surface triplets. Each test set constitutes a mutually exclusive 10% of the dataset used while the remaining 90% are used for training (according to the 90-10 testing scheme, Section 2.4.1)

Set	Mean Propensity of Surface Atoms	StDev of Surface Atoms Propensities	Mean Propensity of Binding Site Atoms	Distribution Shift
Set 1	-0.334	0.058	0.353	11.84
Set 2	-0.344	0.065	0.262	9.32
Set 3	-0.306	0.073	0.227	7.3
Set 4	-0.308	0.074	0.324	8.54
Set 5	-0.28	0.09	0.469	8.32
Set 6	-0.311	0.084	0.292	7.18
Set 7	-0.3	0.055	0.418	13.05
Set 8	-0.299	0.109	0.34	5.86
Set 9	-0.273	0.097	0.467	7.63
Set 10	-0.311	0.072	0.286	8.29

The distributions of average binding site propensities and average surface propensities are graphically compared in Figure 2-21. Figure 2-21B shows that the binding sites have an average propensity greater than the average propensity of all the triplets on the protein surface in 95% of the structures. Those structures for which STP failed to characterize the required binding site correctly were examined. In two cases (1GX5 and 1HYV), the structures were DNA-binding proteins and it is likely that the 'signal' from the DNA binding site was masking the small molecule binding sites used in the test set. In other cases (1IWH, 1MWQ, and 1O7J), STP picked-out the main binding site (the active sites of 1MWQ and 1O7J and the Heme binding site of 1IWH) but failed to pinpoint the test binding site.

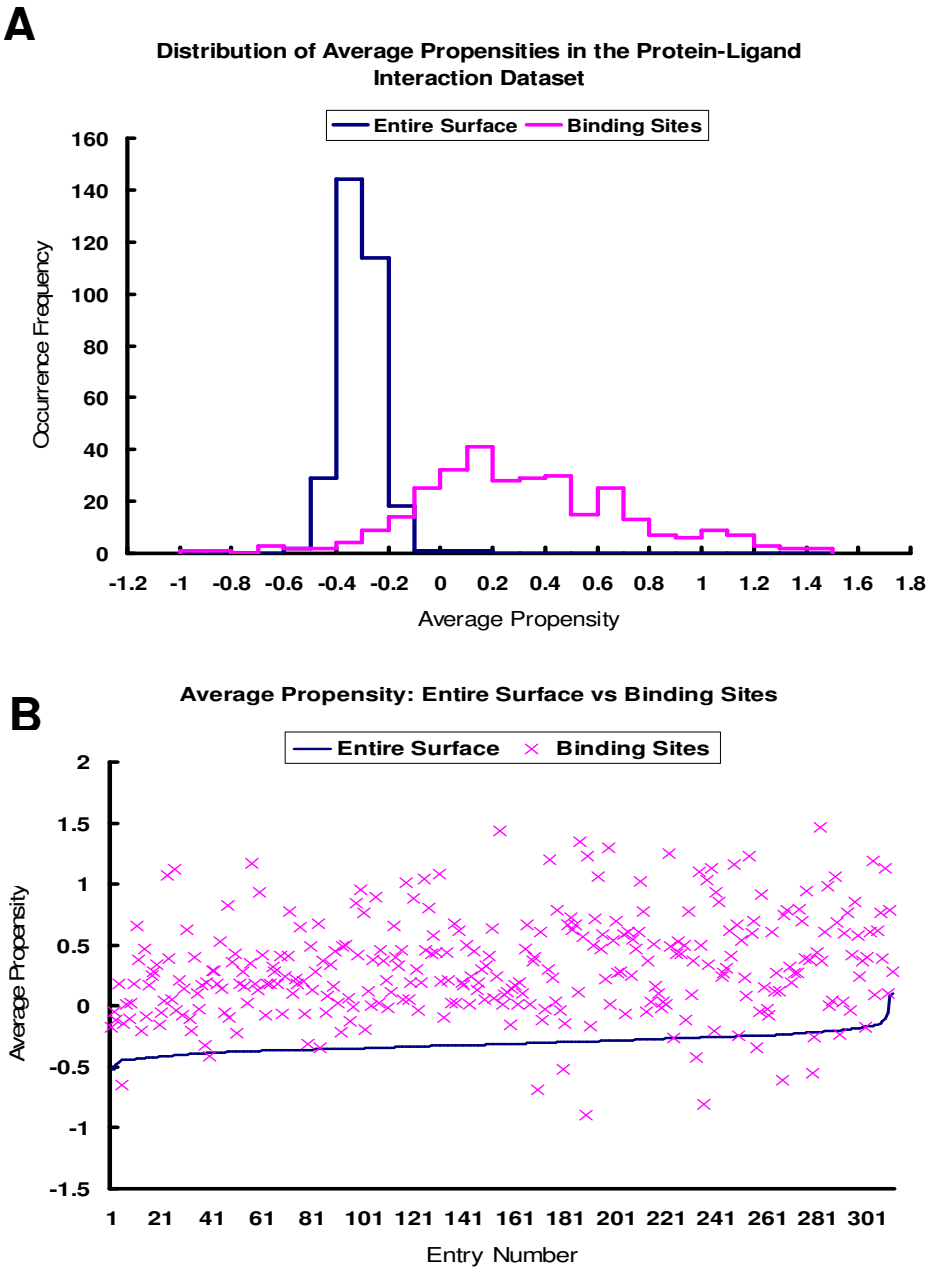

Figure 2-21: Comparison of average propensities of triplets belonging to binding sites and triplets belonging to the entire protein surface in the Protein-Ligand Interaction dataset. Figure A shows how the binding sites distribution is shifted to the right compared to the entire surface distribution while Figure B shows that the average propensity of a binding site is higher than that of the entire surface in approximately 96% of the cases tested.

2.6 Testing the protein and peptide datasets

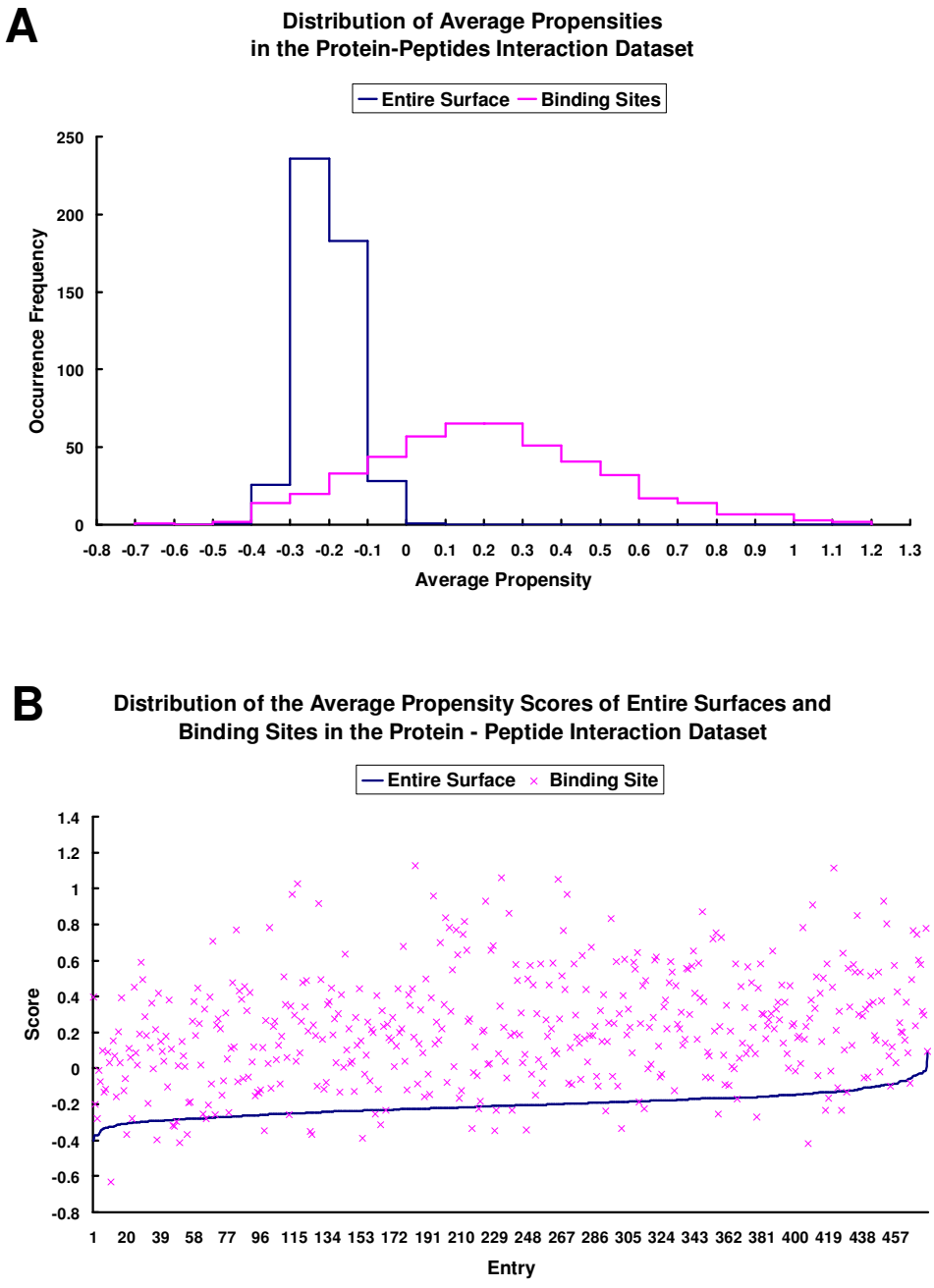

Figure 2-22: Comparison of average propensities of triplets belonging to binding sites and triplets belonging to the entire protein surface in the Protein-Peptide Interaction dataset. Figure A shows how the binding sites distribution is shifted to the right compared to the entire surface distribution while Figure B shows that the average propensity of a binding site is higher than that of the entire surface in approximately 92% of the cases tested.

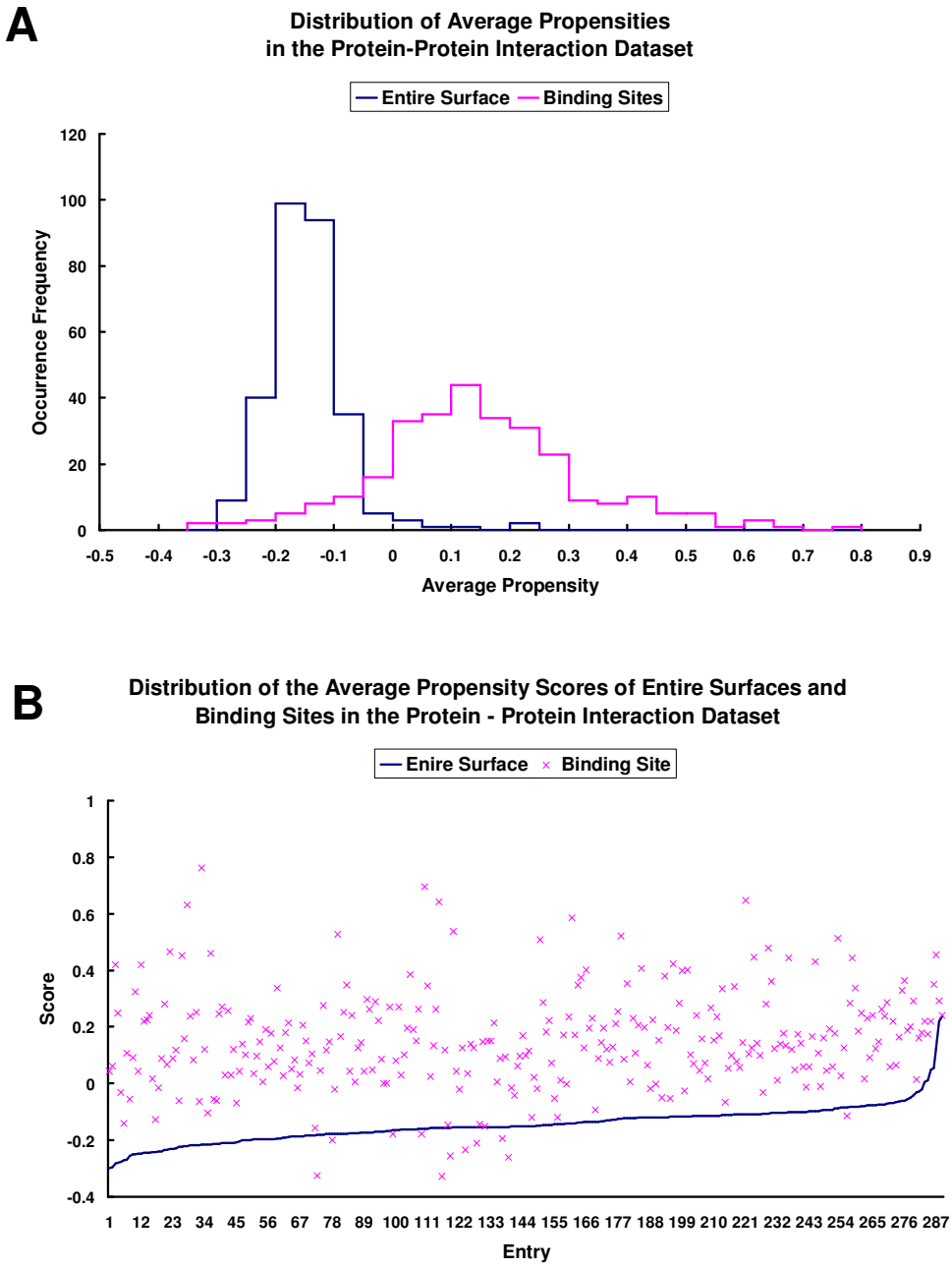

Figure 2-23: Comparison of average propensities of triplets belonging to binding sites and triplets belonging to the entire protein surface in the Protein-Protein Interaction dataset. Figure A shows how the binding sites distribution is shifted to the right compared to the entire surface distribution while Figure B shows that the average propensity of a binding site is higher than that of the entire surface in approximately 96% of the cases tested.

The Protein-Peptide and Protein-Protein score tables were tested by comparing the average propensity of a binding site with the average propensity of all the triplets all over the surface. Similarly to the Protein-Ligand score table, both tests were conducted under the 90-10 scheme (Section 2.4.1). For each score table, each entry is given 2 scores, the first being the average propensity of all binding site triplets and the second being the average propensity of all surface triplets. These 2 attributes are compared (Figure 2-22 and Figure 2-23), showing that binding sites have higher average propensities according to both score tables, indicating the success of the STP algorithm.

When the entire distributions of these 2 attributes are compared (Figure 2-22A and Figure 2-23A), a clear shift to the right is observed for the binding site average propensity distributions for both score tables. Examining these distributions on a case by case basis (Figure 2-22B and Figure 2-23B) shows that 36/475 structures (7.5%) in the Protein-Peptide interaction dataset had their binding sites receive an average propensity less than the average propensity of the entire surface. For the Protein-Protein interaction dataset, 11/289 (3.8%) binding sites received an average propensity less than the average propensity of the entire surface.

The results of this test show that binding sites score higher than the entire surface average. High scoring atoms are therefore colored bright (yellow to red) by the coloring routine, making them identifiable as putative binding sites. With a success rate of 96%, 92%, and 96% for the Protein-Ligand, Protein-Peptide, and Protein-

Protein databases, we conclude that STP is successfully identifying and coloring the binding sites of these interactions.

2.7 Comparison of STP with Other Methods

2.7.1 Comparison with Surfnet

Cavities in protein surfaces are often associated with ligand recognition or enzymatic activity [144-146] and a number of programs (SURFNET [147], Ligsite [148], and PocketFinder [149, 150]) are available to identify such pockets. It is a wide spread belief that the binding site of the protein is located in the largest cleft on the surface [144].

We compare STP's power to locate the binding site with the "largest-pocket" paradigm. We used SURFNET to calculate cavities for the 309 structures in our dataset. The atoms forming the cavities identified by SURFNET were used as input to STP and the cavities were ranked according to the number of high scoring atoms (PatchScore above 70 on a scale of 0 to 100) included. The performance of STP was then assessed by the percentage of cases where the ligand binding site was STP-ranked in the top 1, 2 or 3 cavities. This performance was compared with the occurrence of the binding site in the top 1, 2, or 3 cavities sorted by size only.

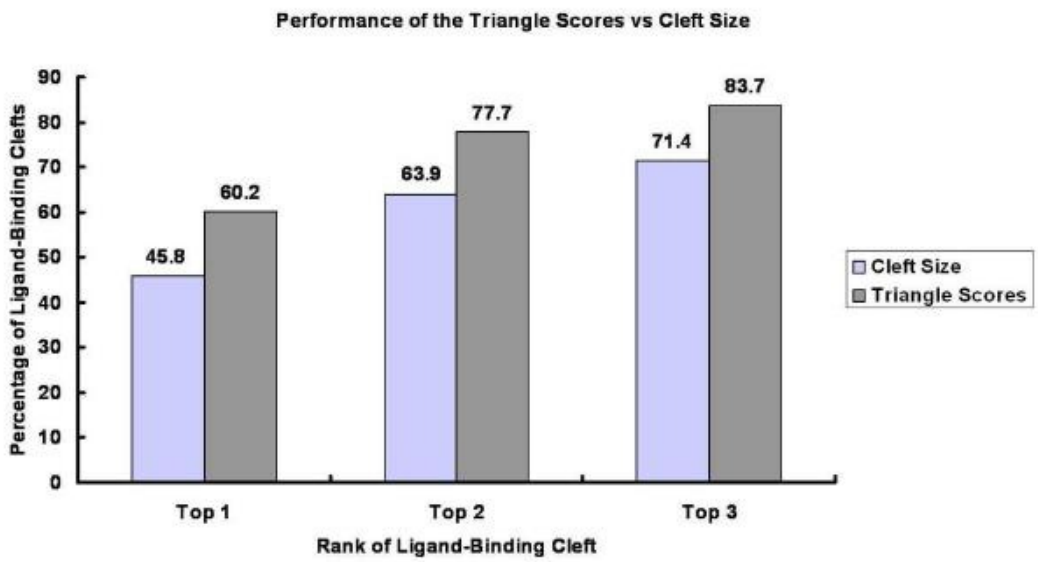

Figure 2-24: The success at predicting binding sites according to cleft size (computed with SURFNET) and STP score of a cleft. Clefts are computed with Surfnet and sorted by volume (cleft size) or by the number of high scoring triplets they include (STP scoring). Graph shows the success rate of predicting the location of binding site by examining the Top 1, 2, or 3 ranked clefts according to each method.

As shown in Figure 2-24 and Table 2-9, STP performs much better than just picking the largest cavity on the surface. On average, Surfnet found 29 cavities on the surface of each protein. 83.7% of the ligand binding sites were discovered in the top 3 cavities ranked by STP. In conclusion, ranking clefts with STP is a better indicator of the location of the binding sites than cavity size and the incorporation of STP with Surfnet as depicted in this experiment gives a better prediction of the binding site than using Surfnet on its own.

Table 2-9: Score and Rank of the experimental binding site each test case according to STP (number of high scoring triplets), and Surfnet (volume). Clefts are computed with Surfnet and sorted by volume (size rank) or by the number of high scoring triplets they include (STP Score).

PDB	STP Score	STP Score Rank	Size Rank	PDB	STP Score	STP Score Rank	Size Rank	PDB	STP Score	STP Score Rank	Size Rank
154L	14	1	1	1JAY	14	1	1	1P0H	8	2	3
16PK	21	1	1	1JBO	24	1	1	1P3D	3	7	1
1AJS	25	2	23	1JG1	4	2	1	1P5Z	14	1	1
1AOE	10	2	1	1JIF	0	4	2	1PI5	18	1	1
1AXW	38	2	2	1JK3	12	4	5	1PIN	16	1	1
1B0U	0	3	2	1JKL	18	1	1	1O6I	75	1	1
1B4P	15	1	1	1JKX	6	3	1	1O6I	57	2	2
1B8O	25	1	3	1JP4	36	1	1	1P1J	9	1	1
1BD0	7	1	13	1JTV	7	1	1	1PJ6	25	2	4
1BKF	27	1	1	1JVP	32	1	1	1PWB	0	10	15
1BUP	5	2	2	1JX4	10	2	2	1Q0N	38	1	1
1BVD	34	1	1	1JZI	0	9	15	1Q0R	18	1	1
1BX4	7	1	5	1K3Y	18	1	1	1Q0R	0	2	15
1BXO	14	1	1	1K4G	12	1	1	1Q1A	19	1	1
1BYQ	8	1	1	1IYH	40	2	1	1Q36	20	2	9
1C1L	10	1	4	1JPZ	31	2	2	1Q4U	12	1	1
1C4Q	7	2	2	1J1N	34	2	3	1Q92	36	1	2
1CCW	0	7	2	1JZ8	17	2	1	1Q9R	7	1	5
1CG6	1	2	2	1KA1	3	3	1	1QD1	15	1	4
1CRU	16	2	2	1KB0	25	2	13	1QGI	28	1	1
1CRU	18	1	3	1KEI	17	1	1	1QH5	29	1	6
1CSH	2	1	1	1KJQ	10	2	1	1QH5	25	2	7
1CZQ	2	2	3	1KLL	0	4	10	1QHO	51	1	3
1D2S	16	1	1	1KM6	3	6	1	1QHO	0	10	28
1D3G	37	1	2	1KMQ	17	1	1	1QJC	36	1	2
1D3G	37	1	2	1KMV	16	1	1	1QJP	5	1	5
1DAD	18	1	1	1KPF	8	1	1	1QK3	14	4	3
1DBW	5	2	4	1KQR	14	1	4	1QNR	47	1	1
1DF7	4	2	1	1KQW	37	1	1	1QPC	4	3	1
1DIM	28	1	2	1KT6	26	1	2	1QV0	60	1	2
1DL2	8	2	3	1KWF	50	1	1	1QXY	30	1	1
1DL2	0	6	25	1L5O	3	3	1	1QZ5	0	3	2
1E19	22	2	2	1L8N	23	1	8	1QZ5	15	1	3
1E2K	34	2	3	1LC3	41	1	1	1QZ5	0	3	11
1E4M	0	7	8	1LJN	9	1	1	1R2Q	10	1	2
1E4M	0	7	12	1LKD	21	1	2	1R5R	41	1	2
1E4M	0	7	22	1LKD	0	2	8	1R6D	27	1	1
1E6W	18	3	1	1LLF	63	2	3	1R6W	18	1	1
1E6Y	53	1	1	1LO7	0	5	5	1R87	46	1	2
1EEX	3	2	4	1LPC	14	1	1	1R8S	6	1	14
1EJ0	5	1	1	1LQT	32	2	2	1QMG	6	5	3
1ELU	53	1	1	1LRI	19	1	1	1PZG	74	1	1
1EU1	22	2	6	1LUQ	23	1	6	1Q74	17	4	5
1EVL	14	2	2	1LXK	33	1	1	1QW9	31	1	1
1EWF	41	1	1	1LZJ	34	1	1	1RA2	8	1	1
1EXM	15	1	8	1M15	34	1	1	1RDQ	10	3	13
1EYN	10	1	1	1M26	13	4	21	1RFF	1	13	5
1F0L	22	1	1	1M26	15	3	27	1RGE	10	3	6
1F2U	27	1	1	1M2K	18	1	2	1RRM	11	2	3

1F5N	8	5	2	1LVW	4	6	6	1RWH	42	1	1
1F6B	18	1	1	1M2R	24	1	1	1RYA	28	2	1
1F74	7	2	3	1M4I	2	2	2	1S1D	14	1	1
1F8E	0	6	1	1M7Y	10	2	16	1S2A	53	1	1
1F9V	2	4	5	1ME4	22	1	1	1SJW	22	1	1
1FCY	21	1	3	1MFA	30	1	6	1SL4	0	2	1
1FCY	2	4	6	1MG5	11	1	1	1SR7	11	1	8
1FK5	13	1	1	1MJH	0	2	3	1STY	16	1	1
1FNC	21	2	1	1MP8	9	2	1	1SU2	2	5	7
1FP2	44	1	1	1MR3	1	4	8	1T46	7	1	1
1FP2	44	1	1	1MRK	16	1	5	1T46	3	4	2
1FRB	41	1	1	1MXG	3	3	11	1TBB	71	1	1
1FTK	11	1	1	1MXI	1	3	2	1TBF	50	1	1
1FZQ	12	1	6	1MZ9	33	3	2	1TH6	8	1	1
1G0O	9	1	5	1N08	0	5	4	1TX4	36	1	1
1G1T	5	1	8	1N1T	21	1	4	1TX4	8	2	23
1G2N	30	1	2	1N2E	24	1	3	1U4B	34	1	1
1G3M	26	2	1	1N3Z	5	3	7	1U4B	1	9	4
1G3M	5	4	8	1N5S	21	1	2	1U4G	16	1	1
1G6H	9	1	7	1N5S	21	1	3	1U7G	4	9	5
1GA2	35	1	3	1N6A	10	2	2	1UDC	22	1	1
1GA2	14	3	10	1N83	24	1	2	1UKV	6	2	2
1GAI	41	1	1	1N8K	24	1	3	1UOG	12	2	2
1GG6	0	4	6	1N8V	61	1	1	1UR1	48	1	1
1GHE	12	2	2	1N9B	9	1	1	1URX	33	1	1
1GKL	6	4	5	1NB9	11	1	2	1US0	47	1	1
1GM7	46	1	1	1NF9	1	4	14	1TAD	27	1	1
1GNX	0	5	8	1NN5	20	1	1	1RXQ	50	2	4
1GOR	35	1	1	1NNF	3	2	1	1URS	53	1	1
1GS5	6	1	1	1M7G	10	3	4	1UTP	10	1	1
1H2B	32	1	3	1M7G	15	2	5	1UU3	16	1	1
1H61	36	1	1	1M7G	9	4	7	1UU6	52	1	1
1H6H	19	1	2	1MV8	10	4	4	1UUY	4	1	1
1H8D	16	1	2	1MV8	5	6	15	1UWC	0	4	6
1HMT	42	1	1	1N62	14	3	11	1UXA	21	1	1
1HNJ	26	1	2	1NRJ	1	6	1	1UY4	14	1	2
1HP1	0	9	25	1NSC	2	6	11	1UYY	2	5	2
1HTW	2	3	27	1NYW	33	1	2	1UYY	0	6	5
1HX0	34	1	2	1O6G	40	1	1	1V0L	31	1	1
1HX0	0	18	11	1O7G	19	1	15	1V2X	2	5	1
1I0V	23	1	1	1O7J	0	15	44	1V3H	20	1	5
1I12	7	1	1	1O7Q	46	1	4	1VHT	0	3	1
1I1N	30	1	1	1O8V	31	1	1	1VHW	28	1	8
1I24	20	1	1	1O97	43	1	2	1VK5	2	2	4
1I3H	5	1	5	1OBD	14	1	1	1VLB	1	4	6
1I4F	10	1	1	1OBD	1	4	2	1W0P	1	4	26
1I58	8	1	1	1OC2	28	2	3	1W3L	13	1	5
1I58	7	2	2	1OD6	12	1	1	1WMS	16	1	1
1G8K	348	1	1	1ODM	51	1	2	2MSB	5	7	20
1I76	14	1	2	1ODZ	50	1	2	2NLR	27	1	1
1ICM	31	1	1	1OE8	12	1	1	2TPS	21	1	6
1ID0	30	1	1	1OFL	12	1	1	3CHB	8	1	20
1IE9	52	1	3	1OFL	9	4	2	3DFR	6	1	1
1IN4	23	1	2	1OGO	14	1	2	3MAN	32	1	1
1IS3	10	1	3	1OH0	29	1	2	3MBP	17	1	1
1IW0	19	2	4	1OI6	12	1	1	3STD	45	2	12
1IWH	31	1	1	1OJJ	45	1	2	4UAG	18	1	1
1IYB	27	1	1	1OJJ	32	2	3	5P21	29	1	1

1J1G	13	1	1	1OQ5	36	1	2	6CEL	21	2	2
1J1M	0	6	5	1OS6	43	1	1	7ATJ	16	1	1
1J54	10	1	1	1OW4	18	1	1	1V00	10	5	38
1JA9	10	1	1	1OWE	36	1	1				

2.7.2 Comparison With Q-SiteFinder and the Method of Morita et al. [151]

A number of binding site prediction methods use GRID-like searches [152] in which interaction energies are calculated between a probe atom and the surface of the protein. We compared the performance of STP against two such approaches; one implemented in the program Q-site finder and the other described in [151]. Both use high scoring probes as seeds for a clustering process that attempts to locate the most energetically favorable locus for a ligand.

A dataset of 35 structurally distinct proteins in the unbound state which share structural similarity with 35 proteins in the ligand-bound form was created by Laurie and Jackson (2005) [153]. This dataset was used to check the performance of the STP on proteins in the unbound state. The unbound proteins were superimposed onto their bound homologues. Ligands were then extracted to mark the binding sites in the unbound proteins. The cavities on these proteins were extracted with SURFNET and then ranked by STP. The binding site of 1 of these structures (1PHD) was an internal binding site and was omitted from the analysis.

Figure 2-25 and Table 2-10 summarize the performance of STP in comparison with Q-SiteFinder, and the method created by [151]. The binding site is located in the top

predicted location by STP in 74% of the cases. This compares with 56% and 76% for the other two methods (Figure 2-25). The binding site is located in the top two predictions in 85% of the cases (the comparable hit rates for the other two approaches are 70% and 82% respectively (Figure 2-25).

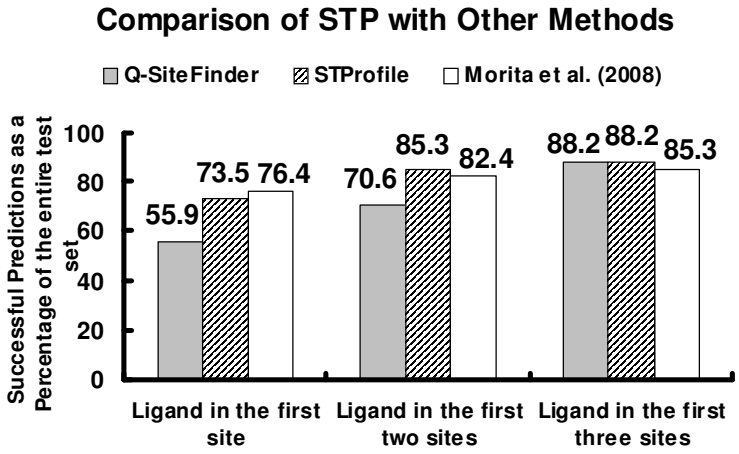

Figure 2-25: Comparison of the performance of STP with Q-SiteFinder and the method by Morita et al. (2008) [151] shows STP to be a competitive and successful binding site prediction program. Surface clefts are extracted with SURFNET and ranked by STP to produce the automated STP prediction of the location of the binding site.

STP succeeds at identifying ligand binding sites in the structures 3APP and 1BYA where both other methods failed [151]. In both these cases the ligands are large and long and this has hindered their prediction: the STP method is independent of the ligand size. We can define a false positive as an experimentally determined binding site which is not among the top 3 predicted STP sites for that protein. There are 4 such structures out of the total 34 in this test data set (1NNA, 1PDY, 1HSI, and 6INS). For 1NNA and 1PDY their heteromerization sites dominated the signal and were predicted over their small molecule ligand binding sites. For 1HSI and 6INS, we can find no documented function for the predicted patches. Thus the false

positive rate from this data set is 2/34 =5.9% (though we cannot exclude the possibility that these predicted patches play an as yet undiscovered role in ligand binding).

Table 2-10: Rank of each test structure according to STP, Q-SiteFinder, and the method by Morita et al. (2008). Items marked with an asterix are greater than 3 (indicating a bad prediction); AVG is the Average Rank of the ligand binding site

ID	STP Rank	Morita et al. (2008) [151] Rank	Q-Site Finder [153] Rank
A6U	1	1	1
1QIF	1	1	1
3APP	1	>10*	4*
1DJB	1	1	2
1BYA	1	>10*	4*
1CGE	1	1	1
1IFB	2	1	1
1HSI	4*	>10*	7*
1A4J	1	1	1
1IME	2	1	1
1NNA	4*	1	2
1AHC	1	1	1
2TGA	1	6*	2
4CA2	1	1	3
1PDY	5*	1	3
1PSN	1	1	1
3LCK	3	1	3
1BRQ	1	1	1
1BBS	1	1	1
1STN	2	1	1
1PTS	1	1	1
2RTA	1	1	1
2CTB	1	4*	1
2CBA	1	1	3
1KRN	1	1	1
2SIL	1	1	3
1L3F	1	1	1
1YPI	2	1	2
1CHG	1	2	2
6INS	4*	2	>10*
2PTN	1	1	1
3P2P	1	1	1
5Cpa	1	3	3
7RAT	1	1	1
AVG	1.56	2.15	2.12

2.8 STPWater

Water plays an important role in protein folding, function, recognition, and interaction [154-156]. It plays an important role in the recognition of Pro-rich ligands by the Abl-Src Homology 3 domain [157] and in the binding of galectin-1 to disaccharide lactose [158]. Moreover, adding water to a biomolecular complex can increase the specificity and affinity of the interaction, leading to various applications in drug design [159]. A new version of STP (STPWater) is designed that incorporates surface water molecules into the triplet patterns. The Protein-Ligand dataset (Section 2.3.1) was used in the creation of this score table. A 15^{th} Atomic Group was added for water: "O2H2". STPWater follows the same methodology of the STP algorithm presented in Section 2.2, and only differs by the incorporation of water molecules with the surface of the protein.

2.8.1 Sampling Useful Water Molecules

The Protein-Ligand Dataset (Section 2.3.1) contains 113726 water molecules out of which 97019 are surface accessible. Using all these molecules in the generation of the surface triplets would lead to the molecular surface of the protein being made up of a lot of water molecules, and the propensity signals of other Atomic Groups might be lost. The large number of water molecules around the protein surface is due to the crystallization procedures which proteins go through, ending up in a space filled with water molecules. Therefore, sampling "useful" water molecules to be assessed by STP is important.

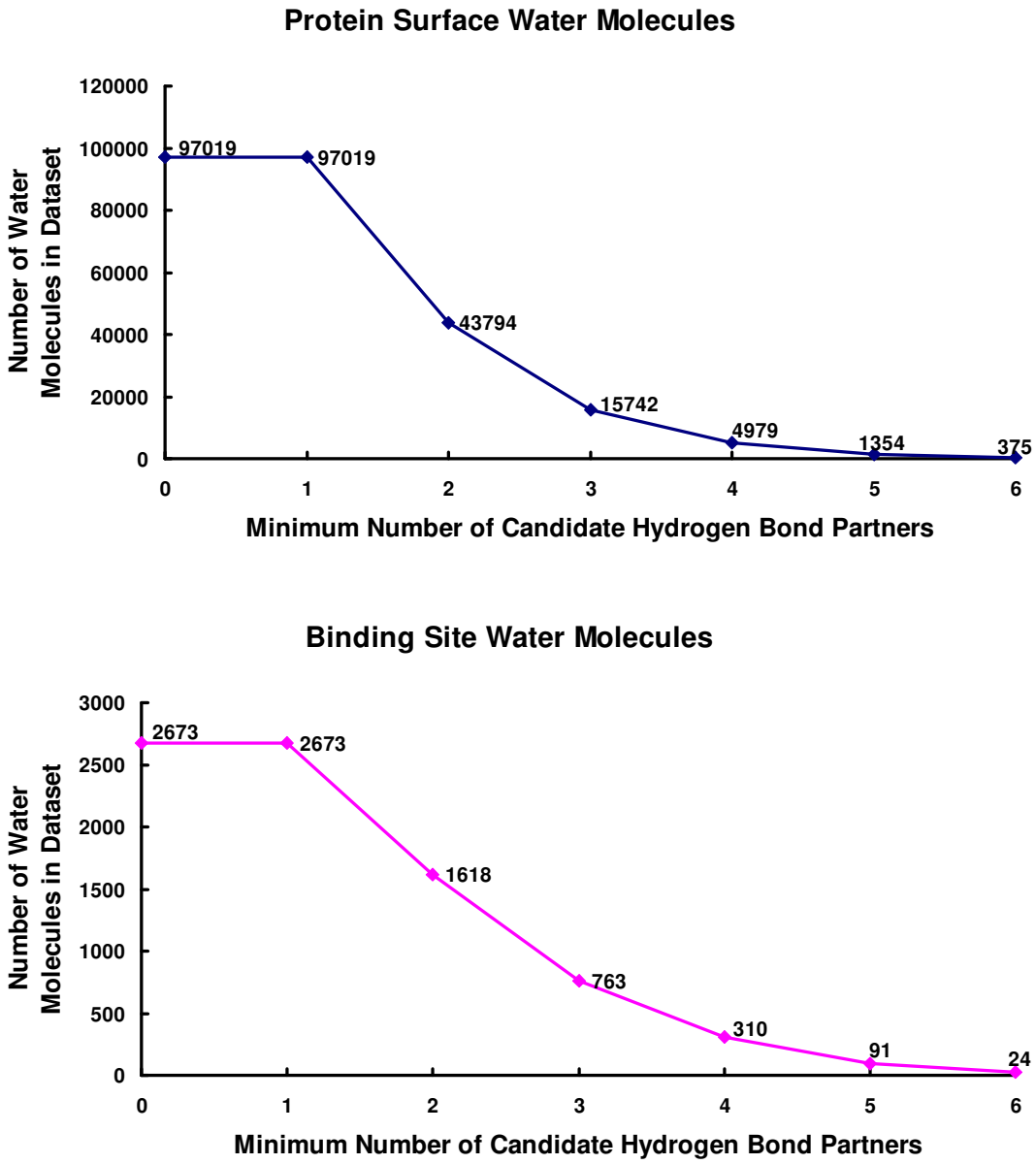

Figure 2-26: The number of water molecules in the Protein – Ligand Dataset (Section 2.3.1) and how it varies with the minimum number of candidate hydrogen bond partners. HBPlus [160] counts the number of possible hydrogen bond interactions for a certain atom based on the distance and bond orientation of this atom and its surrounding partners. An minimum number of hydrogen bond partners = 2 means an atom has at least 2 neighboring atoms capable of forming hydrogen bonds with it.

We define the "usefulness" of a water molecule by the number of possible hydrogen bonds it can make with the protein surface. The hydrogen bonds between water molecules and the protein are calculated by HBPlus [160]. This program studies the

geometries (distance, angles) of different hydrogen bond donor/acceptor pairs to assess their ability to form a hydrogen bond. Water molecules were sampled and the number of possible hydrogen bonds they participate in was calculated (Figure 2-26). Two thresholds were used in assessing the "usefulness" of a water molecule: Threshold A labeling a water molecule as useful if it can participate in at least 3 hydrogen bonds, and Threshold B labeling a water molecule as useful if it can participate in at least 4 hydrogen bonds. These 2 thresholds were chosen since a lot of water molecules can participate in 2 hydrogen bonds and this might lead to over-sampling of these molecules. Restricting the sampling of water molecules to those that can participate in at least 5 hydrogen bonds is also unreasonably high (such situations exist in very special circumstances) and might therefore lead to under-sampling of water molecules. Thresholds A and B are used to create two versions of STPWater, and their performance is tested and compared.

2.8.2 Comparison between the Two Water Thresholds and Regular STP

The two version of STPWater are referred to as STPWater3 for STPWater created with Threshold A and STPWater4 for STPWater created with Threshold B. Both versions have been tested by comparing the distributions of average propensities of triplets belong to binding sites and all surface triplets. The 90-10 cross validation scheme was used again (Section 2.4.1). Each structure in a 10% subset of the dataset was tested with an unbiased score table created from the remaining 90% of structures. The test scheme would calculate 2 attributes for each structure: the average propensity of all binding site triplets and the average propensity of all surface triplets. The comparison of the distributions of these two attributes is detailed in Figure 2-27 and Figure 2-28. The difference between the average propensity of

binding site triplets and that of all surface triplets is quite similar for both *STPWater3* and *STPWater4*. Further tests have been designed to further scrutinize the performance of these 2 versions of STPWater.

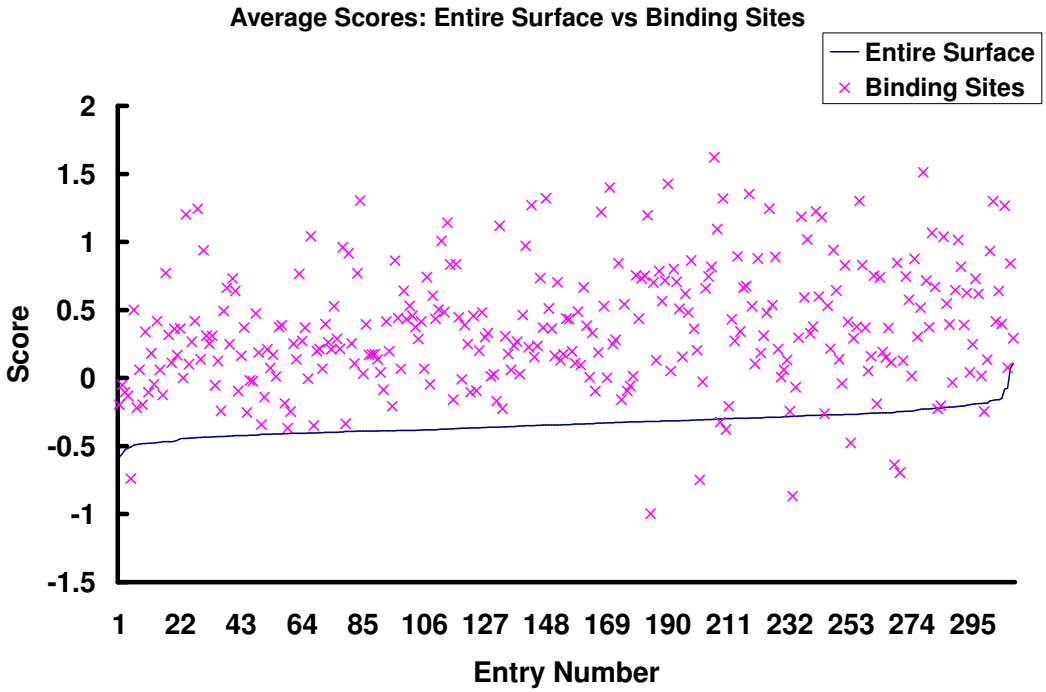

Figure 2-27: Comparison of the Average Propensities of binding sites triplets with the average propensities for all surface triplets as calculated by the STPWater3 score table. 11 structures (3.5%) exhibit binding sites with a lower average propensity than the average propensity of all surface triplets.

STPWater3 and STPWater4 are compared with regular STP. The comparison is done on the basis of two criteria: the number of structures whose average propensity for all binding site triplets is less than the average propensity for all surface triplets and the proportion of "Top Triangles" (Section 2.4.1) that are in the vicinity of a ligand (Figure 2-29 and Figure 2-30). For the second test, each structure is colored with a nonbiased score table (based on the 90-10 testing scheme) and the PatchScores are scaled from 0 to 100. Then "Top Triangles" are detected and checked if they occur at

a maximum distance of 5Å from the ligand of that structure. The proportion of those triplets that are within this distance was calculated and used to compare the performance of the 3 versions of STP.

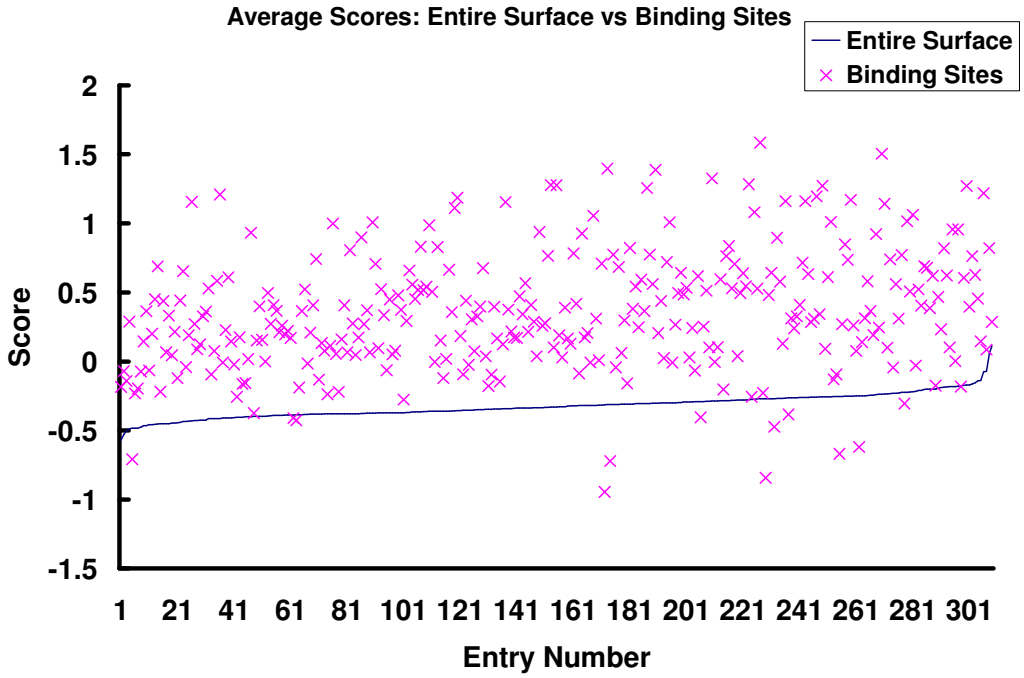

Figure 2-28: Comparison of the average propensity scores of binding sites triplets with the average propensities for all surface triplets as calculated by the STPWater4 score table. 13 structures (4.2%) exhibit binding sites with a lower average propensity than the average propensity of all surface triplets.

STPWater3 shows an advantage over STPWater4 and regular STP when it comes to the number of structures with average propensities of binding sites triplets compared to the average propensities of all triplets on the surfaces of those structures (Figure 2-29): 11 for STPWater3, 13 for each of the original STP and STPWater4. Comparing the proportion of "Top Triangles" (Section 2.4.1) that are within the vicinity of a ligand shows that STPWater3 and STPWater4 have very similar performances; both of them outperforming regular STP (Figure 2-30). It is therefore

concluded that STPWater is a good extension of STP and has a better performance. The difference between STPWater3 and STPWater4 is minimal, with STPWater3 performing slightly better (Figure 2-29).

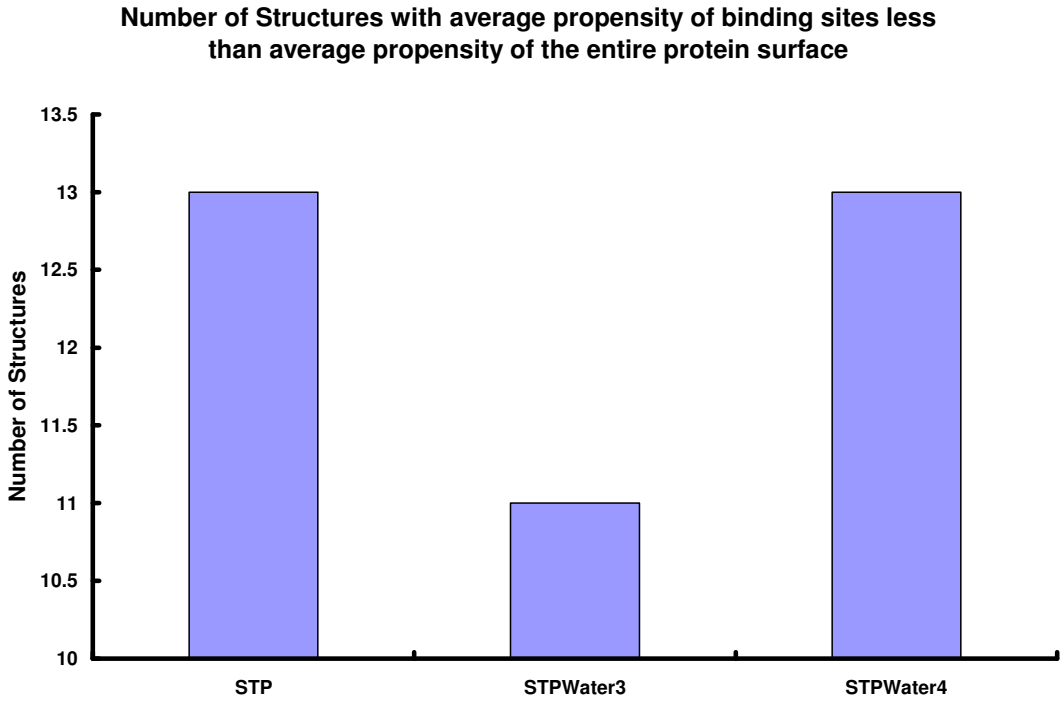

Figure 2-29: The number of structures that have binding sites with a lower average propensity than the average propensity of all surface triplets according to the original STP, STPWater3, and STPWater4 score tables.

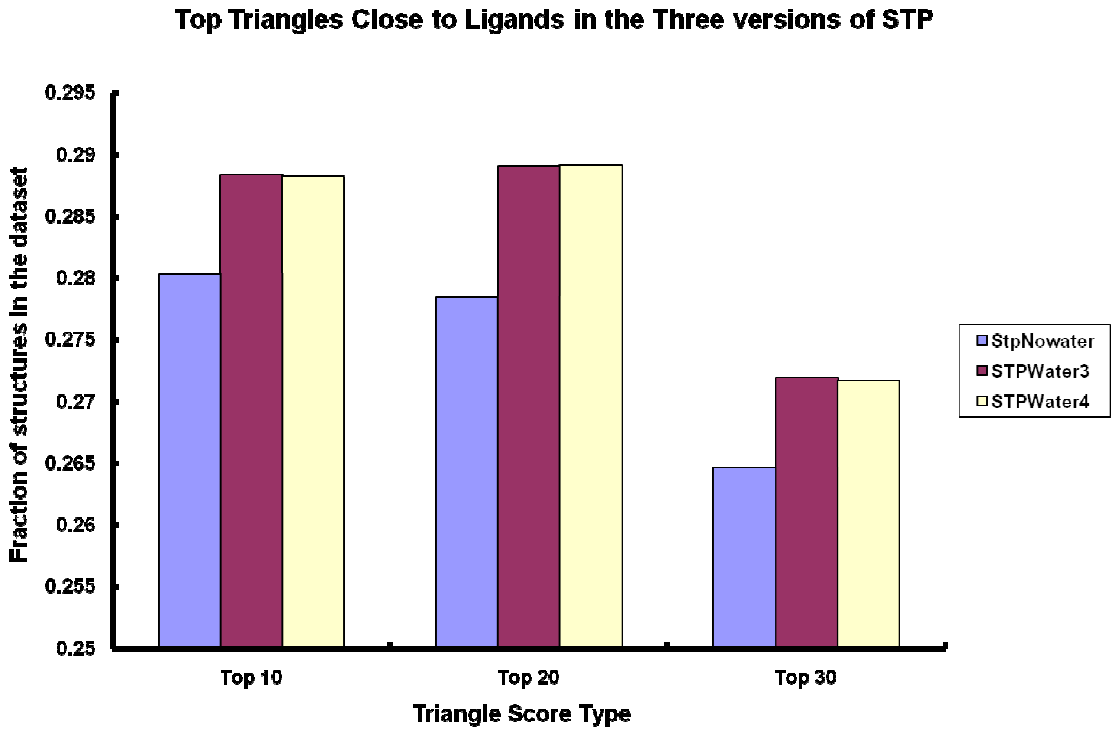

Figure 2-30: The fraction of high scoring triplets that are in the vicinity (5) of a ligand according to the STP, STPWater3, and STPWater4 score tables. High scoring triplets are those with PatchScores higher than 90 (Top 10), 80 (Top 20), or 70 (Top 30) on a scale of 0 to 100. Results show a minute improvement by incorporating water molecules into the STP prediction algorithm. STPWater3 and STPWater4 correspond to STP versions including water molecules with at least 3 or 4 candidate hydrogen bond partners respectively.

3 Applications of the STP Propensities

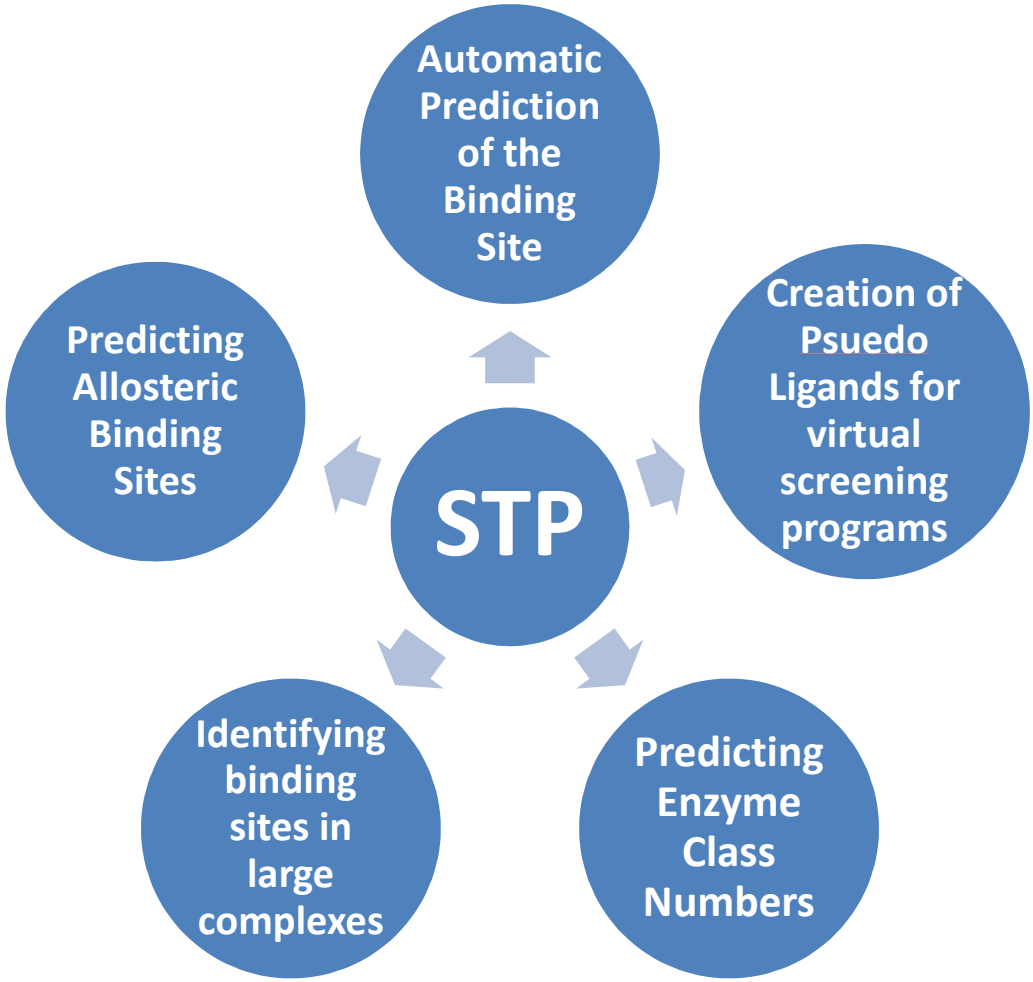

Figure 3-1: Various applications of the STP propensity scores in automated prediction of binding sites, pseudo-ligand creation, E.C. class number prediction, allosteric binding sites, and binding surfaces in large multi component complexes.

Various applications of the STP propensity scores are discussed in this chapter (Figure 3-1). The STP algorithm explained in Chapter 2 relies on color coding the surface which is then processed by the human eye. Section 3.1 details an STP-dependent scoring function that scores the likelihood of a specific point around the protein to be adjacent to a binding site. This is then exploited to create an automatic version of STP that predicts a binding site without any manual interpretation or

intervention. The same routine is also used to create pseudo-ligands, used as templates for docking programs, locating the area to be tested in docking. In Section 3.2, the ability to distinguish the E.C. numbers of enzymes by their surface triplet is tested. Six propensity score tables are created, each corresponding to the likelihood of STP triplets to exist in the active sites of a certain class of enzymes. These score tables are then used to predict the E.C. numbers of unknown active sites. Section 3.3 details the use of STP propensities in re-ranking docking orientations, using STP as a scoring function that filters orientations generated by docking programs. Finally, section 3.4 explains the application of the STP coloring routine to large multi-component complexes while section 3.5 details the capability of STP to predict the location of allosteric binding sites.

3.1 Automated Identification of Protein Binding Surfaces

Energy based programs are widely used for various applications like docking [69, 71, 161], molecular dynamics [162], predicting binding sites [151-153]. We use the STP propensities to create an STP-based energy-like function, capable of scoring points in space around the protein based on their distance from surface triplets, and the propensity of these triplets. High scoring points (referred to as STP site points) will then be flagged as favorable by STP since they will be close to high scoring atomic groups and triplets. This leads to two direct applications. First, STP site points can be used for automated prediction of the binding site, without any human intervention (like looking at a colored surface). This is useful in case a user wants to predict the locations of binding sites on many PDB structures and does not have enough time to manually look at the colored surface of these structures. Second, STP site points will serve as pseudo-ligands for docking programs when the ligand is not available. Many

docking programs [69, 71, 161] rely on the existence of a known ligand to specify the area in which compounds would be docked on a certain receptor. In the absence of a ligand, these programs are given mock points which may or may not be accurate. STP site points are binding site predictions and will provide more accurate pseudo-ligands for these docking programs.

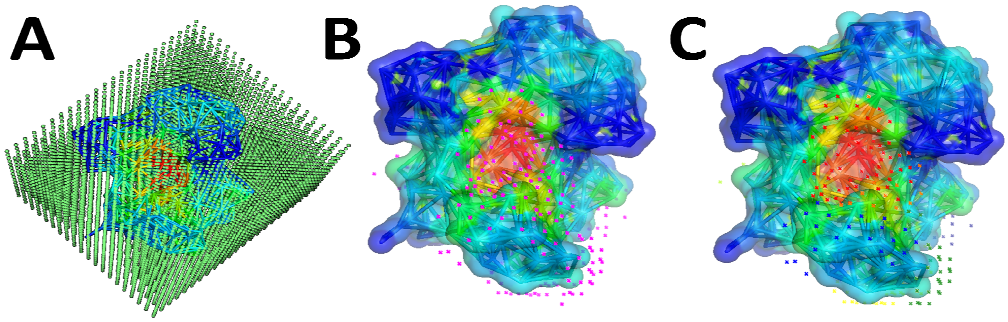

Figure 3-2: A cube of grid points around the protein is created (A). These grid points will be scored via a scoring function. Some grid points will be discarded due to distance or score parameters (B). The remaining grid points will be clustered into groups, as shown by the different colors of gridpoints in (C). Each cluster will have a score, indicating the likelihood of that cluster to be the binding site.

A forcefield is defined as a scoring function capable of scoring an "interaction value" at any point in space. Points in space are created (as a grid) and scored via a scoring function (Figure 3-2). High scoring points are identified as STP site points. Adjacent site points are clustered to form an STP site point cluster; leading to the generation of several clusters serving as suggestions for the location of the binding site. This poses several questions:

1. How will the grid points be generated?
2. What scoring function will be used to assign scores to different points?

3. How will points with high scores be identified (i.e. what is a high score)?
4. How will the points be clustered?

3.1.1 Grid Point Generation

A grid was defined as a rectangular box around the protein (Figure 3-2 A). The grid generation program searches for maximum and minimum coordinates in the protein structure and pads 4Å in each dimension (x, y, and z). Grid points are then created at a resolution of 2Å (i.e. we have a grid point at a step distance of 2Å). For each entry in the representative dataset (Table 2-2), grid points around the protein surface were classified as *binding site points* if the minimum distance to any known-ligand atom is less than or equal to 2Å. Grid points that are more than 4Å away from the protein (*maximum distance threshold*) were neglected, as well as points closer than 2.5 Å (*minimum distance threshold*). The minimum distance threshold (2.5Å) was chosen to minimize steric clashes between site points and the protein atoms. The maximum distance threshold (4 Å) was chosen to avoid extending the binding site far into the space around the protein (Figure 3-2 B). Once the scoring function has been created, this maximum distance threshold will be tested and tuned.

3.1.2 The Scoring Function and Identification of STP Site points

The *Site pointScore* of any point (x, y, z) in space depends on the STP propensity of the protein surface triplets, and the distance between this point in space and these triplets (Figure 3-3). Therefore, a general energy-mimicking formula for the scoring function was designed as by Equation 3-1. Site pointScores of a certain structure are often scaled onto a 0-1 interval, creating a *scaled Site pointScore*. We define the notion of a *score-band*, similar to the notion of Top Triangles in Section 2.4.1. The

score-bands used in this chapter and their corresponding scaled Site pointScores are: Top15 (0.85 to 1), Top20 (0.8 to 1), Top25 (0.75 to 1), Top30 (0.7 to 1), Top35 (0.65 to 1), Top40 (0.6 to 1), Top45 (0.55 to 1), Top50 (0.5 to 1), and Top60 (0.4 to 1).

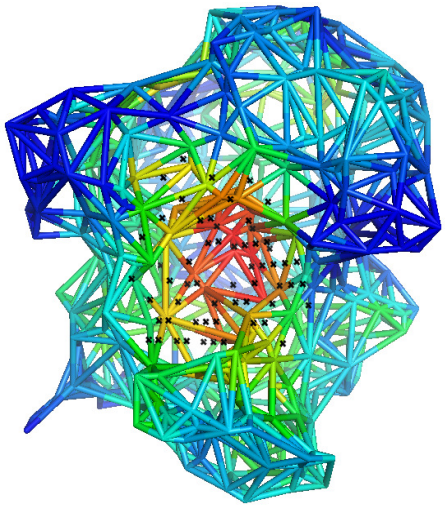

Figure 3-3: Each point in space (a selection is shown in black) will be scored based on the surface triplets of the protein, and the distance between that point and those triplets (distance between the point and the centroid of a triplet is used). General form of the scoring function is given in Equation 3-1.

$$SitePointScore = \sum_{i=1}^{N} \frac{\Pr opensity_i}{d_i^{\alpha}}$$

Equation 3-1: The score given to any point in space is based on the propensities of all the protein surface triplets, and the distance "d" between the centroids of these triplets and this point in space.

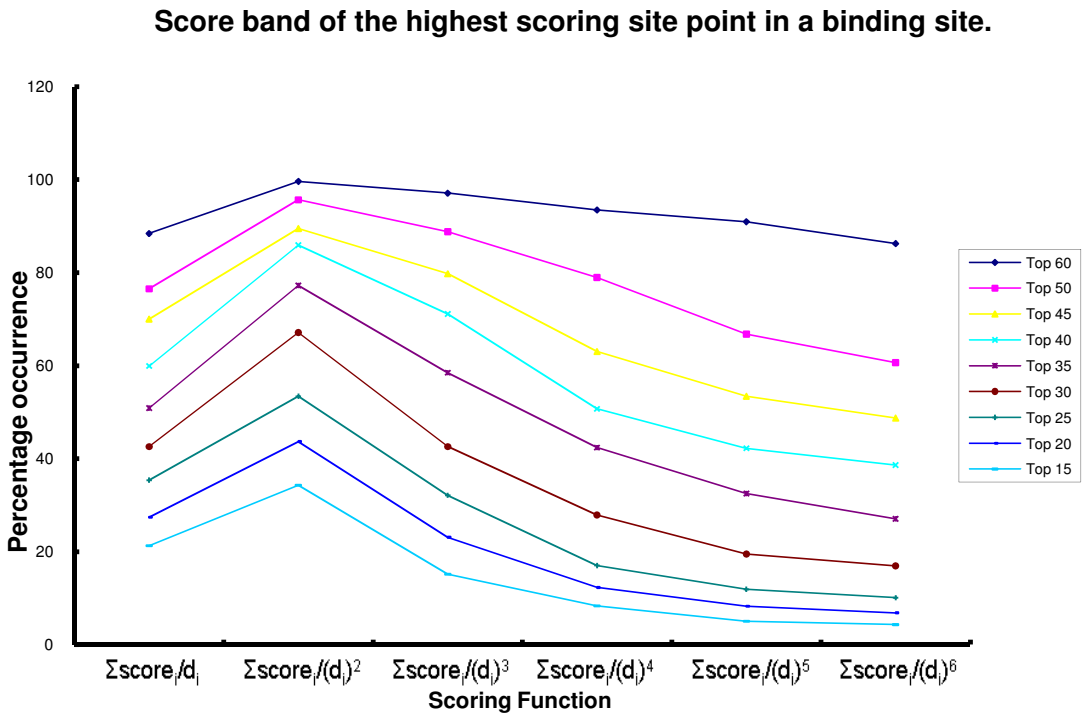

Figure 3-4: The assessment of several scoring functions (listed on the x-axis of the Figure). For each structure, grid points around the protein are scored with the various functions and scaled 0 to 1. The grid points that belong to the binding site (computed by a maximum 2 Å distance from the ligand atoms) are identified. The highest scoring point in a binding site is checked for its score band (Section 3.1.2). Score bands tested are: Top15 (0.85 to 1), Top20 (0.8 to 1), Top25 (0.75 to 1), Top30 (0.7 to 1), Top35 (0.65 to 1), Top40 (0.6 to 1), Top45 (0.55 to 1), Top50 (0.5 to 1), and Top60 (0.4 to 1). Figure shows that for all score bands, using $\Sigma score/d^2$ leads to binding site points having higher scores.

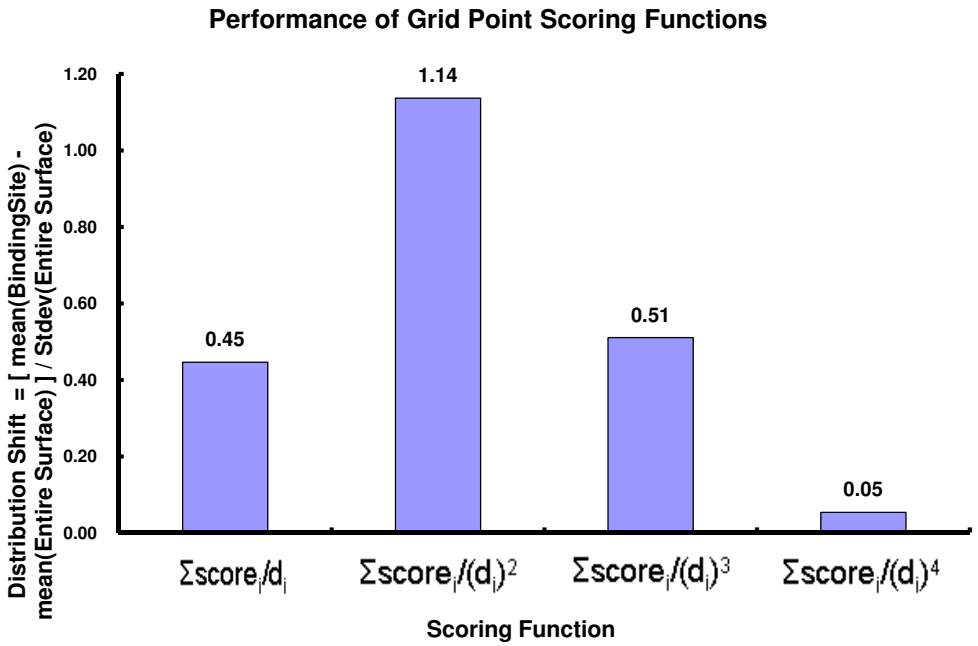

Figure 3-5: The distribution shift between the average Site pointScore of a binding site point and that of the entire protein surface site points. The higher the shift, the more distinguished are the binding site points from the rest of the site points around the surface. The x-axis denotes the power to which the distance is raised, according to Equation 3-1.

Six values were nominated for α in Equation 3-1 (1 through 6) and tested. Grid points were scored according to each function, then scores where normalized on a 0 to 1 scale. Each structure in the protein-ligand dataset was tested to check whether it has at least 1 binding site point in each of the score bands Top 60, Top 50, Top 45, Top 40, Top 35, Top 30, Top 25, Top 20, and Top 15 (Section 3.1.2); where a Top N band represents points with scores in the top N% of the scoring range (so a Top 60 band refers to scores 0.4 to 1). The maximum distance threshold is tested and tuned. Grid points around the protein were scored and normalized, and then the minimum distance to the protein from any grid point in the Top 10 band was calculated. These distances had an average of 2.29Å and Standard deviation 0.46Å and are listed in Table 3-1. The maximum value was measured for entry 1USO (3.53Å). Therefore

the maximum distance threshold was fixed to 4Å (giving around 0.5Å margin over the maximum distance calculated).

The proportion of structures with points in these bands where recorded for the six nominated scoring functions and are presented in Figure 3-4. At "α = 2", binding site points are scored within higher score-bands than with the other tested values for α (Figure 3-4). This *site point score* assesses the likelihood of a point in space belonging to a binding site. The distance of a point to a certain triplet is important and plays a role in determining the role of this point in space. Scaling the distance by a power of two gives the most suitable distance contribution to the scoring function. Another remarkable result from this graph is that at "α = 2", 99.63% of the test cases had binding site points in the Top 60 band. Consequently, a first screening phase will be added to this functionality, which removes all points that do not lie in the Top 60 band (Figure 3-2 B). Figure 3-5 shows the distribution shift of the scores of binding site gridpoints minus the scores of all the gridpoints around a protein. The results backup the findings in Figure 3-4.

Table 3-1: Minimum distance from a Top 10 grid point to the protein

PDBID	Distance (Å)	PDBID	Distance (Å)	PDBID	Distance (Å)	PDBID	Distance (Å)
1UU6	1.45	1OC2	1.95	1J1M	2.23	3MBP	2.60
1ME4	1.50	1P3D	1.96	1UY4	2.23	1FTK	2.61
1EU1	1.50	1HX0	1.96	1FZQ	2.24	1MR3	2.64
1V2X	1.51	2NLR	1.96	1UKV	2.24	1M26	2.66
1MXG	1.53	1U4G	1.96	1E4M	2.24	1AJS	2.66
1DF7	1.53	1OQ5	1.97	1G6H	2.25	2MSB	2.66
1RRM	1.56	1I24	1.97	1HP1	2.25	1K3Y	2.67
1SR7	1.59	1UUY	1.97	1R6W	2.25	1JZ8	2.67
1O6I	1.60	1M7Y	1.97	1W3L	2.27	1SJW	2.68
1GA2	1.61	1LPC	1.98	1HMT	2.27	1GHE	2.68
1Q4U	1.61	1FK5	1.98	1JKL	2.28	1D2S	2.69
1QW9	1.61	1Q92	1.99	1MJH	2.28	1QH5	2.70
1RA2	1.62	1EEX	1.99	1KMV	2.28	1B0U	2.70
1F9V	1.62	1TX4	2.00	1OFL	2.28	1L5O	2.71

1O6G	1.64	3STD	2.01	1LLF	2.28	1ODM	2.73
1WMS	1.66	1N8V	2.01	1EVL	2.29	1OGO	2.73
1C1L	1.66	1T46	2.01	1P1J	2.29	1LRI	2.74
1KQR	1.66	1IW0	2.02	1IS3	2.29	1P5Z	2.74
1RXQ	1.67	1SU2	2.02	1LVW	2.30	1EJ0	2.75
1M2K	1.67	1JX4	2.02	1KMQ	2.30	1NN5	2.76
1LQT	1.68	1RWH	2.05	1RGE	2.31	1O7J	2.78
1QGI	1.69	1MXI	2.05	3DFR	2.31	1FP2	2.79
1O7Q	1.69	1B4P	2.05	1F0L	2.31	1I4F	2.80
1KPF	1.70	1VHW	2.06	1KM6	2.32	1O8V	2.80
1E2K	1.71	1GS5	2.06	1I58	2.32	1M7G	2.81
1AOE	1.71	1H2B	2.06	2TPS	2.32	1FNC	2.81
1UYY	1.72	1JZI	2.06	1TBB	2.33	1QK3	2.82
1BYQ	1.73	1IWH	2.07	1R6D	2.33	1F5N	2.82
1I0V	1.73	1TBF	2.07	1M2R	2.34	1LUQ	2.84
1GKL	1.73	1JA9	2.07	1GG6	2.34	1JP4	2.86
1P0H	1.73	1QD1	2.08	4UAG	2.35	1TAD	2.87
1GAI	1.74	1MP8	2.09	1STY	2.36	1MFA	2.88
1J1G	1.74	1B8O	2.09	1NRJ	2.37	1Q0R	2.89
1BXO	1.74	1DL2	2.09	1F6B	2.37	1JKX	2.91
1MWQ	1.75	1D3G	2.09	1QPC	2.37	1SL4	2.91
1HTW	1.75	5P21	2.10	1QXY	2.38	1BUP	2.92
1UXA	1.76	1IYH	2.10	1I76	2.39	1N62	2.92
1M15	1.77	1VHT	2.10	1G2N	2.39	1PIN	2.92
1Q36	1.77	1M4I	2.11	1JPZ	2.39	1F2U	2.93
1Q0N	1.78	1AXW	2.11	1KIC	2.39	1V0L	2.93
1PI5	1.78	1R5R	2.11	1QNR	2.40	1S1D	2.94
1GNX	1.78	1NB9	2.12	1JBO	2.41	1RFF	2.94
1HNJ	1.78	1F74	2.12	1UOG	2.41	1UR1	2.94
1I3H	1.79	1ELU	2.13	1KLL	2.41	1U7G	2.95
1Q9R	1.79	1LO7	2.13	1K4G	2.42	1VLB	2.97
1MRK	1.79	1CCW	2.13	1H61	2.42	3MAN	2.97
1MV8	1.81	1CZQ	2.13	1KA1	2.43	16PK	2.98
1ICM	1.82	1UWC	2.14	1KJQ	2.43	1OWE	2.99
1TH6	1.83	1CG6	2.14	1UTP	2.44	1ODZ	2.99
3CHB	1.83	1N1T	2.14	1FCY	2.44	1OS6	3.01
1OJJ	1.85	1O97	2.15	1I12	2.44	1DAD	3.05
1H4E	1.85	1NYW	2.16	1QV0	2.44	1N6A	3.05
1G0O	1.85	1MZ9	2.16	1NF9	2.45	1R87	3.06
1V00	1.85	1G8K	2.16	1GM7	2.45	1LJN	3.06
1BKF	1.86	1Q1A	2.16	1PZG	2.45	1FRB	3.06
1ID0	1.86	1N9B	2.16	1S2A	2.45	1EXM	3.07
1RYA	1.87	1I1N	2.16	1UU3	2.45	1OW4	3.07
1LC3	1.87	1N8K	2.17	1N5S	2.46	1PJ6	3.09
1G3M	1.87	1JTV	2.17	1O7G	2.46	1J54	3.10
1R2Q	1.87	1KB0	2.17	1BX4	2.47	1N83	3.10
1IE9	1.89	1JVP	2.17	1G1T	2.47	1L8N	3.11
1MG5	1.90	1JIF	2.17	1NNF	2.47	1KEI	3.11
1PWB	1.90	1W0P	2.18	1KQW	2.49	1LZJ	3.13
1DIM	1.91	1N3Z	2.18	1GX5	2.50	7ATJ	3.17
1OE8	1.91	1IYB	2.18	1BVD	2.51	1URX	3.18
1NSC	1.91	1JG1	2.18	1E19	2.52	1V3H	3.20
1H6H	1.92	1CSH	2.18	1E6W	2.52	1LKD	3.24
1BD0	1.92	1QMG	2.19	1LXK	2.53	1HFU	3.26
1DBW	1.92	1HYV	2.19	1CRU	2.54	1OBD	3.26
1N2E	1.93	1QJC	2.19	1J1N	2.54	1JK3	3.26
1QHO	1.93	1E6Y	2.20	1GOR	2.54	1EWF	3.33
1EN2	1.93	1VK5	2.20	1JAY	2.56	1IN4	3.33

1RDQ	1.94	154L	2.21	1U4B	2.56	1QJP	3.42
1EYN	1.94	1N08	2.21	1OH0	2.57	6CEL	3.44
1R8S	1.94	1H8D	2.21	1C4Q	2.57	1US0	3.53
1Q74	1.94	1KT6	2.21	1URS	2.57		
1F8E	1.94	1UDC	2.22	1OD6	2.58		
1KWF	1.95	1QZ5	2.22	1OI6	2.60		

3.1.3 Clustering STP Site points

After the distance threshold and the scoring functions have been tuned, the only question left to answer is which clustering method should be used. Clustering points according to their spatial and score distributions is a computationally hard problem. There are many known clustering methods algorithms. Some methods apply pure computational clustering without any heuristics (QT clustering [163]) while others rely on calculating probabilities for each point to be in a certain clustering (EM clustering [164], Bayesian Clustering [165]). The use of various other methods that utilize machine learning (SVMs [166], Neural Networks [167], and Bayesian machines [168]) have also been mentioned in the literature. The QT clustering algorithm was chosen to be used in this work because of its simple methodology and its independence from heuristics and probability calculations which might complicate the task. The pseudocode of the original QTClustering algorithm is found in Algorithm 3-1.

Algorithm 3-1: QT Clustering Algorithm Pseudocode

```
QT_Clust(G, d)
        If diameter (G) ≤ 7
                Output G #originally was if |G| ≤ 1 but adapted to suit site point generation
        Else
                For all i ε G
                        Set flag ← True
                        Set A_i ← {i} #A_i is the cluster started by i
                        While flag = True and A_i ≠ G
                                Find j ε (G - A_i) such that diameter (A_i U {j}) is minimum
                                If diameter (A_i U {j}) > d}
                                        Set flag ← False
                                Else
                                        Set A_i ← A_i U {j} #add j to cluster A_i
                                End If
                        End While
                End For
                Identify set C ε {A_1, A_2, … , A_i} with maximum cardinality
                Output C
                Call QT_Clust(G-C, d)
        End If
End QT_Clust
```

The QT Clustering algorithm uses a "maximum diameter" to define a cluster. The *diameter* of a set of points is defined as the maximum distance between any 2 points belonging to that set. Consequently, the diameter of a set containing the binding site points for each entry in the dataset was measured. To recap, *binding site points* are grid points around the protein surface. These points were classified as if the minimum distance to any known-ligand atom is less than or equal to 2Å (Section 3.1.1). Only entries with one ligand were taken into consideration. Diameters of these site point sets ranged between 7Å and 25.97 Å, with a mean of 14.7Å and standard deviation of 3.8Å (Figure 3-6). These values were used to define the geometric limits of the site point clusters, keeping these clusters spatially similar to the binding site points sets observed in the training set. Hence, a cluster is defined to have a maximum diameter of 15Å and a minimum diameter of 7Å.

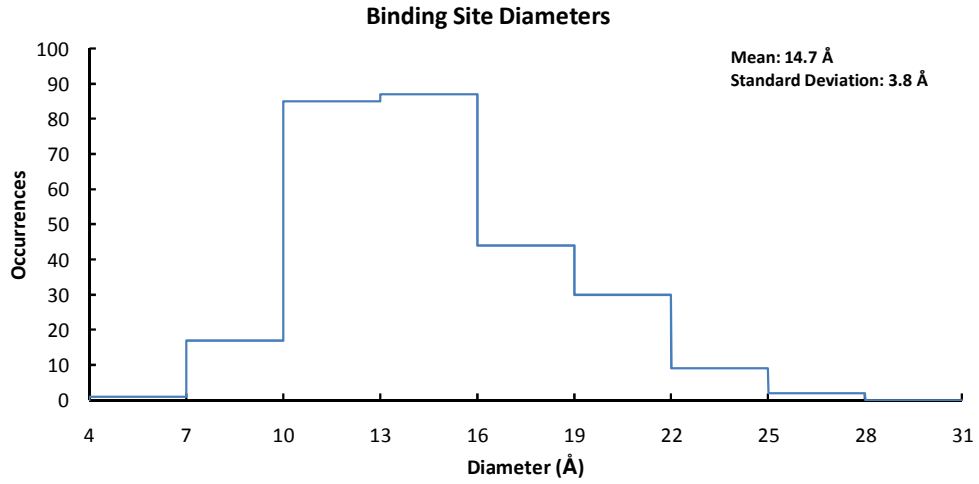

Figure 3-6: The distribution of diameters of binding sites in protein-ligand dataset. Only entries with one ligand were considered. A diameter of a set of points is defined as the maximum distance between any two points in that set of points. A set of binding site points is defined as all grid points within a maximum distance of 2 Å from a ligand atom.

Table 3-2: The Diameter of a binding site, defined as the maximum distance between any 2 site points belonging to that binding site (Section 3.1.1). A histogram is presented in Figure 3-6.

Entry	Diameter (Å)	Entry	Diameter (Å)	Entry	Diameter (Å)	Entry	Diameter (Å)
1CZQ	7.00	1JTV	11.71	1N2E	14.35	1BKF	16.92
1FTK	7.27	1MFA	11.71	1RA2	14.35	1EJ0	16.92
1G0O	7.54	1PIN	11.71	1G6H	14.41	1EXM	16.92
1JA9	7.54	3STD	11.71	1GM7	14.41	1VK5	17.09
1O7G	7.67	1FNC	11.96	1S2A	14.41	1H6H	17.15
1GKL	8.16	1RGE	11.96	1SL4	14.41	1GS5	17.26
1PWB	8.28	1HMT	12.04	1EVL	14.48	1BVD	17.37
1E2K	8.96	1IS3	12.04	3MBP	14.48	5P21	17.37
1V0L	9.07	1MXI	12.04	1KJQ	14.62	1TX4	17.43
1MWQ	9.29	1ODZ	12.04	1KMV	14.62	1NRJ	17.54
1MRK	9.50	1L8N	12.12	1ODM	14.62	1ME4	17.60
1O7J	9.50	1LJN	12.12	1PI5	14.62	1N62	17.60
1UTP	9.50	1QNR	12.12	1DF7	14.68	1EN2	17.76
1GG6	9.80	1RXQ	12.20	1RFF	14.68	1R6D	17.76
1Q36	9.80	1GOR	12.28	1STY	14.68	1URX	17.93
1VHW	9.80	1LPC	12.28	1I76	14.88	1F2U	17.98
1LUQ	10.00	1QK3	12.29	1N08	14.88	1G1T	17.98
1O97	10.00	1MP8	12.36	1N8V	14.88	1IWH	18.04
1I4F	10.19	1GX5	12.60	1F5N	14.95	1JPZ	18.04
1NNF	10.19	1OH0	12.60	1HTW	14.95	1MXG	18.31
1TH6	10.19	1QW9	12.60	1ICM	14.95	1MG5	18.41

1KB0	10.29	1SR7	12.60	1F9V	15.01	1N83	18.73
1N1T	10.29	1LZJ	12.68	1OS6	15.01	1RRM	18.73
1Q92	10.29	1JVP	12.75	1MJH	15.08	1VLB	18.89
1QXY	10.29	1LVW	12.75	1S1D	15.14	1M2K	18.99
1EYN	10.48	1N6A	12.75	1UR1	15.14	1P0H	18.99
1I0V	10.48	1H4E	12.83	1URS	15.14	4UAG	18.99
1I3H	10.48	1KM6	12.83	3MAN	15.14	1P1J	19.04
1KQR	10.48	1L5O	12.98	1IN4	15.21	1QMG	19.09
1M2R	10.48	1OGO	12.98	1K3Y	15.21	1Q1A	19.50
1N3Z	10.48	1Q9R	12.98	1OE8	15.21	1GAI	19.65
1NSC	10.48	1SU2	12.98	1PJ6	15.21	1JBO	19.70
2MSB	10.48	1US0	12.98	3CHB	15.21	1G8K	19.85
1AJS	10.57	1WMS	12.98	2NLR	15.34	1LRI	19.95
1GNX	10.57	1C4Q	13.13	1E6Y	15.40	1V3H	19.95
1OW4	10.57	1LC3	13.13	1OD6	15.40	1MZ9	20.09
1V00	10.57	1JX4	13.21	1BUP	15.46	1EEX	20.24
1OI6	10.66	1KLL	13.21	1HP1	15.46	1N8K	20.38
1E6W	10.75	1O8V	13.21	1I24	15.46	1BXO	20.43
1IW0	10.75	1Q0N	13.21	1OWE	15.46	1EU1	20.81
1J1G	10.75	2TPS	13.21	1UDC	15.46	1QD1	20.81
1KEI	10.75	1M15	13.28	1IYH	15.65	1LLF	21.09
1W0P	10.75	1O6G	13.28	1UOG	15.65	6CEL	21.14
1IYB	10.93	1OQ5	13.28	1NB9	15.72	1CSH	21.19
1R5R	10.93	1Q74	13.28	1TAD	15.72	1H2B	21.19
1AOE	11.02	1SJW	13.28	1U4G	15.72	1PZG	21.23
1B8O	11.02	1V2X	13.28	1UKV	15.72	1I12	21.37
1BD0	11.02	1NF9	13.36	3DFR	15.72	1J1N	21.37
1D2S	11.02	1NYW	13.50	1F6B	15.90	1LO7	21.37
1DBW	11.02	1O7Q	13.50	1H8D	15.90	1QJP	21.37
1HYV	11.02	1BYQ	13.57	1KT6	15.90	1GHE	21.42
1J1M	11.02	1H61	13.57	1MR3	15.96	1UY4	21.51
1JP4	11.02	1KA1	13.57	1B4P	16.02	1RWH	21.55
1R6W	11.02	1QPC	13.57	1IE9	16.02	1UU6	21.60
7ATJ	11.02	1UU3	13.57	1W3L	16.02	16PK	21.87
1F74	11.29	1QJC	13.72	1QV0	16.08	1FK5	21.91
1F8E	11.29	1DAD	13.86	1B0U	16.15	1JAY	21.91
1JZI	11.29	1ID0	13.86	1JG1	16.15	1LXK	22.14
1BX4	11.37	1R8S	13.86	1I1N	16.21	1T46	22.49
1CG6	11.37	1RYA	13.86	1KMQ	16.21	1Q4U	22.57
1DIM	11.37	1U7G	13.93	1UUY	16.21	1JKX	22.75
1TBB	11.37	1E19	14.07	1R2Q	16.45	1G2N	23.00
1KIC	11.46	1KPF	14.07	1KQW	16.51	1HNJ	23.22
1UWC	11.54	1RDQ	14.07	1F0L	16.62	1JIF	23.88
1C1L	11.63	1TBF	14.07	1QGI	16.62	1LQT	24.25
1ELU	11.63	1JK3	14.28	1FZQ	16.86	1CCW	24.29
1J54	11.63	1P5Z	14.28	1JKL	16.86	1KWF	25.51
1K4G	11.63	1FRB	14.35	154L	16.92	1EWF	25.97
1UXA	11.63	1M7Y	14.35	1AXW	16.92		

3.1.4 Creating Pseudo-Ligands with STP

STP site points can be used as pseudo-ligands to be input for docking programs like LIDAEUS [69], Autodock [71], and Autodock Vina [161]. These programs depend on a ligand coordinates file to locate the part of the receptor on which the docking partners will be docked. In standard cases, a ligand exists in the crystal structure and is used by these programs. In the absence of a known ligand, the user has to create a pseudo-ligand to be used as a beacon for locating the binding site. While experienced users rely on professional programs like COOT [169] to create such pseudo ligands, inexperienced users try to introduce atoms around the protein. Some of these atoms are either too close or too far from the protein surface. Moreover, there are no guidelines for how big or small a ligand should be. STP provides a simple and useful solution for this problem, by creating pseudo-ligands around a protein.

The primary gain from using STP to create pseudo-ligands for a protein is that STP will use the propensity of surface atoms to suggest where the binding site is, and then attempts to create clusters of atoms around the protein with the highest possible score. Grid points around the protein are generated (Section 3.1.1) and a scaled Site pointScore (Section 3.1.2) is generated per grid point. According to Figure 3-4, 99.63% of the tested cases showed that the binding site includes points with a scaled Site pointScore of at least 0.4 (Section 3.1.2). Therefore, all grid points with scores less than 0.4 are neglected. The remaining grid points are clustered (Section 3.1.3) into sets, called clusters. Every cluster gets a *cluster score* equal to the average

scaled Site pointScore for all site points in that cluster. A mol2[9] file is generated and each grid point represented by a carbon atom. Atoms of the same cluster are grouped in residues. An average user can rely on Pymol [43] to select a residue and extract it to be used a pseudo-ligand. The success of this routine was measured as the percentage of dataset structures whose top 3 ranked pseudo ligands overlap the experimentally verified ligand (overlap measured as an experimental ligand with at least 1 atom within 2 Å of a pseudo ligand). The top ranked pseudo ligand overlaps with the original ligand in 41%, compared with 62% and 66% for the top two and top three ranked pseudo ligands (Figure 3-7). Two examples are shown Figure 3-8.

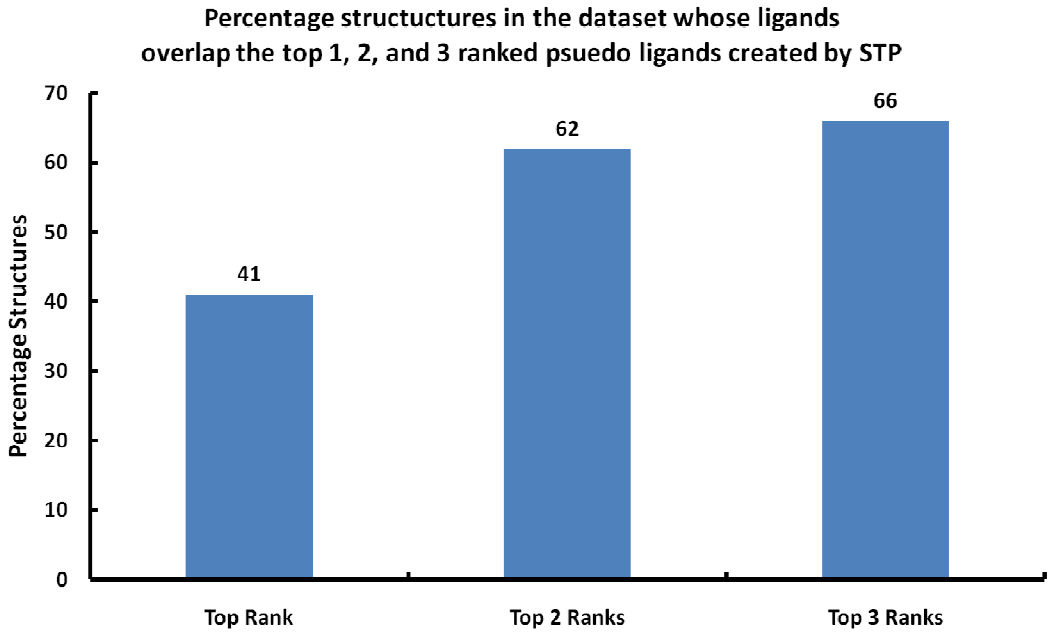

Figure 3-7: The percentage of dataset structures whose experimentally verified ligands overlap with the top 3 ranked STP generated pseudo ligands.

[9] http://tripos.com/tripos_resources/fileroot/pdfs/mol2_format2.pdf

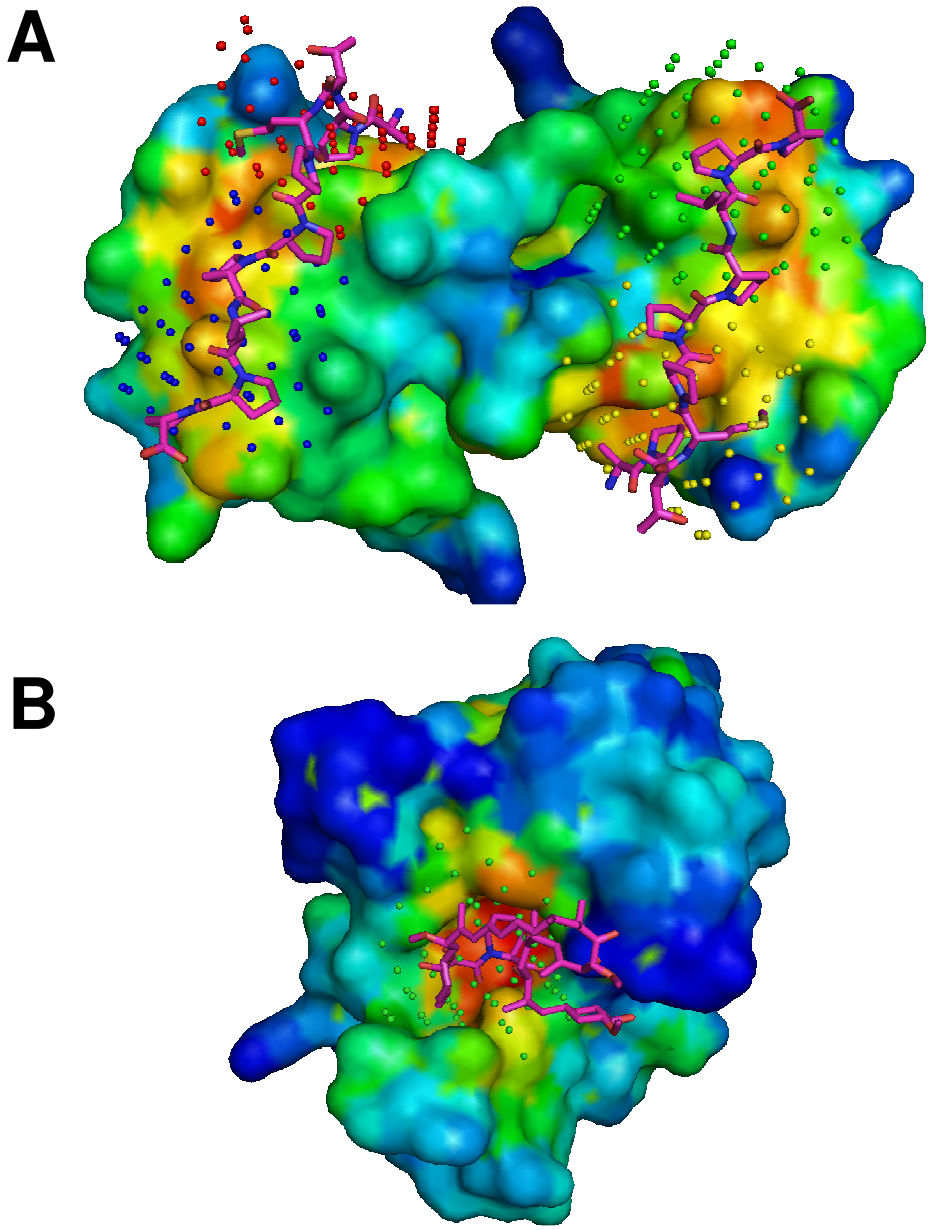

Figure 3-8: ABL Tyr Kinase bound to the synthetic peptides (A) and FKBP12 bound to FK506 (B). Each structure is colored with STP to locate the binding site. Small spheres around the ligand represent the top ranked clusters of STP site points created by STP to serve as pseudo-ligands. Four clusters are presented in A and are colored blue, green, yellow, red (in descending order of average cluster score), while one cluster is shown in B, colored green.

The pseudo-ligand generation program leads to the incorporation of STP with the in-house virtual screening pipeline (Figure 3-10). This automatic system is demonstrated on the *Leishmania Mexicana* CRK3 (Cyclin Related Kinase 3) protein kinase. *Leishmania Mexicana* CRKs are homologues of human Cyclin Dependant Kinases (CDKs). CRK3 is a cdc2-related protein kinase with activity towards histone H1. It has been proven to be essential to *Leishmania mexicana* since disrupting both alleles of the CRK3 gene resulted in changes in cell ploidy [170]. There is no existing structure for CRK3 or its human homologue CDK1 (54% sequence identity). CRK3 has been modelled using modeller [171] based on the structure of the protein CDK2 (58% sequence identity). This model of CRK3 is then targeted with the in house virtual screening pipeline, in a search for possible inhibitors for the active site of CRK3, or for CRK3-Cyclin 6 (CYC6) interaction.

The CRK3 model was studied by STP for the nomination of possible binding sites and the generation of pseudo ligands (Figure 3-10). Two of these pseudo ligands were picked out: the first (referred to as ligand location A), in a ridge edged by Arg'25-Val`28 from one side, and Lys`53-Arg`55 from the other side, and the second referred to as ligand location B, in a deep pocket located around Phe`99-Ala`104 (Figure 3-9).

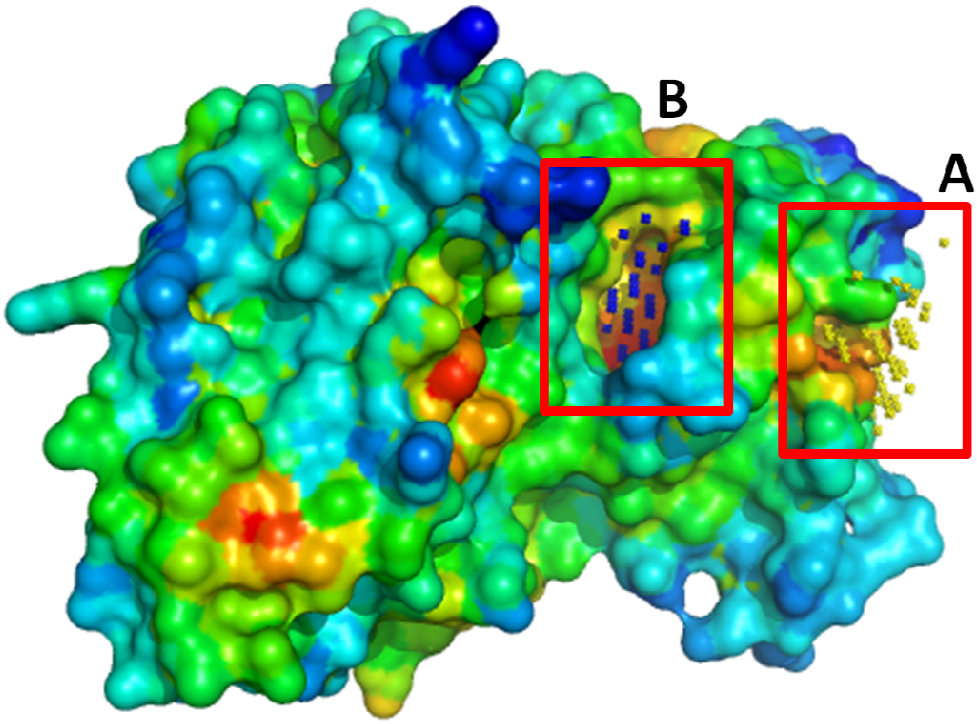

Figure 3-9: The pseudo ligands created by STP on the CRK3 model, indicating ligand location A and ligand location B used in this experiment.

Possible inhibitors were screened with LIDAEUS [69]. The screened compounds were all selected from the Maybridge dataset in the EDULISS database, summing up to 64,000 compounds. The top 250 LIDAEUS hits were then taken and redocked with the higher resolution docking program AutoDock [71]. Highscoring ligands were picked from the autodock experiments, and similar compounds were mined from EDULISS with the program UFSRAT [98]. As shown in Table 3-3, the 250 compounds screened with AutoDock at positions A and B of the CRK3 structure showed high binding affinities (several low nanoMolar compounds). The UFSRAT generated compounds showed an enrichment in the predicting binding affinities, whether with the introduction of predicted picoMolar binding compounds (ligand

location A), or several nanoMolar to microMolar Binding compounds (ligand location B). Figure 3-11 shows the binding of the best two ligands to the STP predicted binding sites.

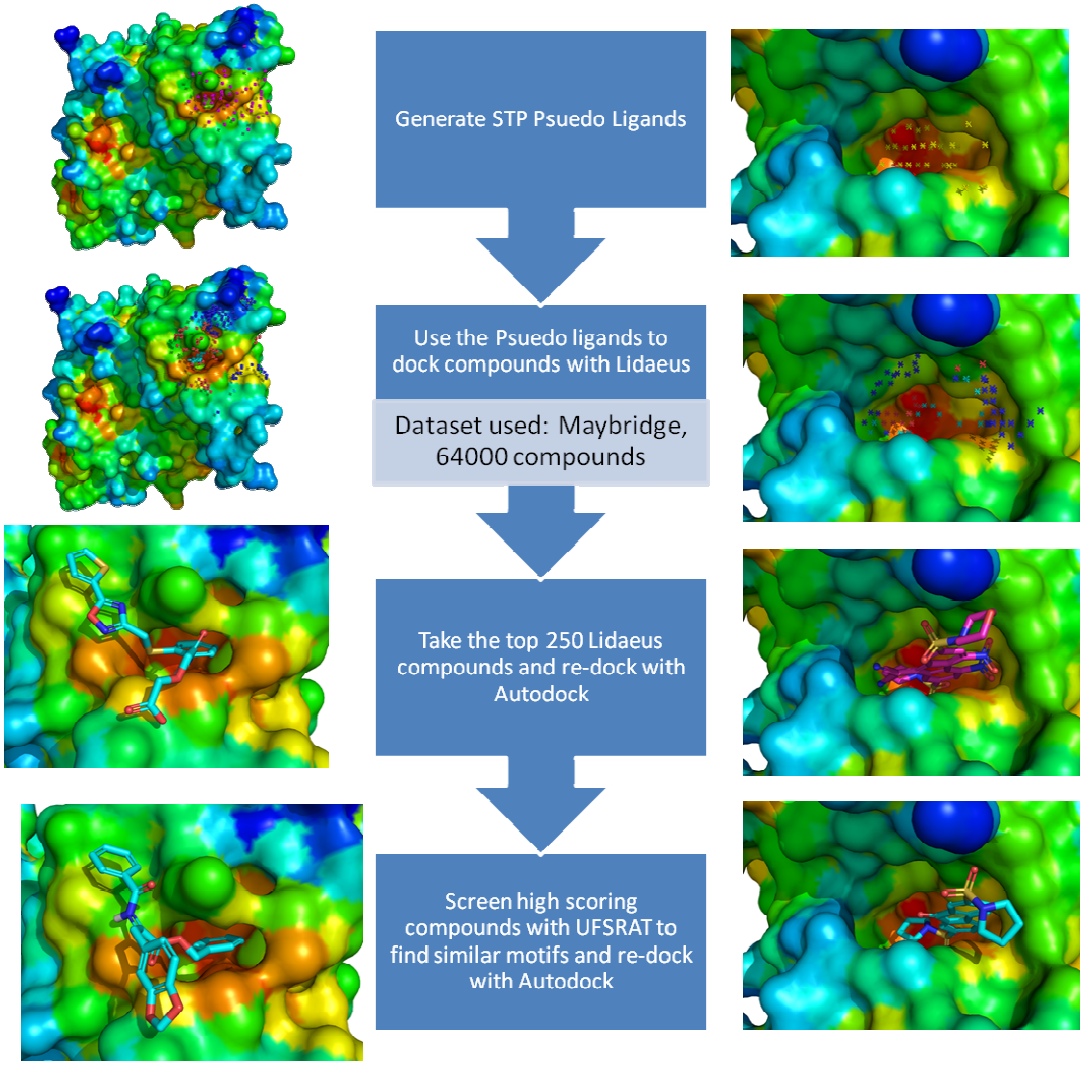

Figure 3-10: The incorporation of STP pseudo ligand creator in the in house virtual screening pipeline generates possible binding partners for CRK3 at 2 different locations predicted by STP.

Table 3-3: The binding affinity and binding energy ranges generated by the AutoDock screenings described in Figure 3-10 indicates a high likelihood of possible inhibitors for the CRK3 binding in two distinct locations predicted by STP.

Virtual Screening Experiment	Affinity Range	Energy Range (kcal/mol)
CRK3 Pseudo Ligand A	[3.52 nM, 45.76 mM]	[-11.53, -1.86]
CRK3 Pseudo Ligand A UFSRAT Homologues	[399 pM, 1.74 mM]	[-12.82, -3.76]
CRK3 Pseudo Ligand B	[725.69 pM, 2.92mM]	[-12.47, -3.46]
CRK3 Pseudo Ligand B UFSRAT Homologues	[22.7 nM, 169.69µM]	[-10.44, -5.14]

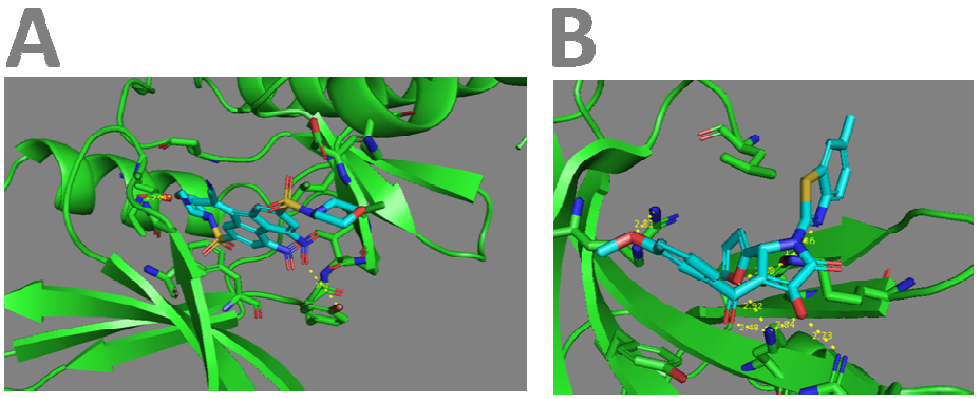

Figure 3-11: The binding of top compounds to the CRK3. A shows the binding of compound 8SPH1-253-022 to CRK3 (position of pseudo ligand 1) at a predicted affinity of 399 pM. B shows the binding of compound 24SPH1-009-712 to CRK3 (position of pseudo ligand 2) at a predicted affinity of 726 pM.

3.2 Predicting Enzyme Class with STP

Proteins interact with their binding partners as a part of larger pathways where each step has a designated function. Since surface atoms and residues play a vital role in these interactions, it is accepted that certain motifs of these atoms will help determine the function binding pocket of a certain protein. Determining the function of a protein is usually approached from sequence and structure homology[27, 66,

172-177] or pattern matching [48, 51, 52, 178-182] perspectives. This section addresses the question of whether STP surface triplets are helpful in predicting the first level E.C. numbers of enzymes. Enzymes are abundant in structural databases, important for molecular function, and well characterized in general (chemically and biologically). A large amount of research is invested in enzymes and that leads to a wealth of information about their active sites and mechanisms through data sources like the CAZymes Analysis Toolkit [183], the CoFactor database [184], and MacIE [185]. The objective is to design a system that takes in a binding site, and predict its enzyme class number based on the triplets it contains.

3.2.1 Enzyme Classes

Enzymes are classified into six major classes and are known by their Enzyme Commission (E.C.) number. An E.C. number of an enzyme is made up of 4 parts, each part indicating a sub-classification of the class indicated by the previous part. In this experiment, we will be interested in the first-level classification of enzymes, signaled by the first digit of the E.C. Number. Oxidoreductases (E.C. 1) are responsible for oxidation or reduction reactions in the body. Transferases (E.C. 2) are responsible for the transfer of chemical groups from substrates to products. Hydrolases (E.C. 3) cleave bonds via hydrolysis. Lyases (E.C. 4) eliminate rings and double bonds, but not through hydrolysis or oxidation. Isomerases (E.C. 5) are responsible for inducing geometric changes in molecules. Ligases (E.C. 6) are responsible for joining molecules together, usually coupled with the hydrolysis of a chemical group on one of these molecules.

3.2.2 The Training Dataset

A representative dataset of the enzymes in the Protein Databank is needed by STP to predict the enzyme class of a protein. The Protein-Ligand Dataset (Section 2.3.1) is used for this purpose; all enzyme structures are extracted and annotated with their Enzyme Class specification. 198 of the dataset's 309 structures are enzymes, distributed among 40 oxidoreductases (E.C. 1), 57 transferases (E.C. 2), 79 hydrolases (E.C. 3), 15 lyases (E.C. 4), 9 isomerases (E.C. 5), and zero ligases (E.C. 6) (Table 3-4). Because our dataset contained no ligases, a different dataset should be compiled to cover the entire span of the Enzyme Classes.

MacIE [185] is an online database of enzymes in the PDB, containing a dataset 260 PDB codes, representing all 249 unique E.C. Numbers and 331 CATH Codes [19]. This dataset is unique by chemical action; every enzyme in this dataset corresponds to a unique chemical mechanism. Unfortunately, using this database proved to be unfeasible. Fourteen out of the first forty analyzed structures had a poorly defined binding site; the ligands (if present) are often defined as one or two metal atoms (for example, PDB structures 1EZV, 1VNC, 2FRV, and 1A7U). This makes it hard to define the binding surface of the protein by STP, which needs a considerably larger ligand to define a binding surface. Using MacIE's enzyme dataset was therefore cancelled. Instead, we added the 13 ligase structures from MacIE to the protein-ligand dataset, creating a comprehensive dataset covering all 6 enzyme classes. Two of these 13 structures were neglected (2TS1 and 1F7U, both have poorly defined active sites), and we end up with a dataset of 209 enzymes (Table 3-4).

Table 3-4: The enzymes in the protein-ligand dataset. Table shows 209 structures distributed among 79 Hydrolases (Enzyme Commission Number "E.C." 3), 9 Isomerases (E.C. 5), 14 Lyases (E.C. 4), 39 Oxidoreductases (E.C. 1), 57 Transferases (E.C. 2). The remaining 11 Ligases (E.C. 6) have been imported from MacIE [185] as the protein-ligand dataset included zero ligases.

PDB ID	Enzyme Class	PDB ID	Enzyme Class	PDB ID	Enzyme Class	PDB ID	Enzyme Class
154L	Hydrolase	1QNR	Hydrolase	1E6Y	Oxidoreductase	1J54	Transferase
1BUP	Hydrolase	1QW9	Hydrolase	1EU1	Oxidoreductase	1JG1	Transferase
1BXO	Hydrolase	1QXY	Hydrolase	1FNC	Oxidoreductase	1JKL	Transferase
1DL2	Hydrolase	1R87	Hydrolase	1FRB	Oxidoreductase	1JKX	Transferase
1E4M	Hydrolase	1RFF	Hydrolase	1G0O	Oxidoreductase	1JVP	Transferase
1F8E	Hydrolase	1RGE	Hydrolase	1G8K	Oxidoreductase	1JX4	Transferase
1GA2	Hydrolase	1RYA	Hydrolase	1H2B	Oxidoreductase	1K3Y	Transferase
1GAI	Hydrolase	1S1D	Hydrolase	1HFU	Oxidoreductase	1K4G	Transferase
1GG6	Hydrolase	1STY	Hydrolase	1IW0	Oxidoreductase	1KJQ	Transferase
1GKL	Hydrolase	1SU2	Hydrolase	1JA9	Oxidoreductase	1L5O	Transferase
1GM7	Hydrolase	1TBB	Hydrolase	1JPZ	Oxidoreductase	1LVW	Transferase
1GNX	Hydrolase	1TBF	Hydrolase	1JTV	Oxidoreductase	1LZJ	Transferase
1GOR	Hydrolase	1TH6	Hydrolase	1KB0	Oxidoreductase	1M15	Transferase
1HP1	Hydrolase	1U4G	Hydrolase	1KMV	Oxidoreductase	1M2R	Transferase
1HX0	Hydrolase	1UR1	Hydrolase	1LC3	Oxidoreductase	1M4I	Transferase
1I0V	Hydrolase	1URX	Hydrolase	1LKD	Oxidoreductase	1M7G	Transferase
1I76	Hydrolase	1UTP	Hydrolase	1LQT	Oxidoreductase	1MP8	Transferase
1IYB	Hydrolase	1UU6	Hydrolase	1MG5	Oxidoreductase	1MXI	Transferase
1J1G	Hydrolase	1UWC	Hydrolase	1MV8	Oxidoreductase	1N08	Transferase
1J1M	Hydrolase	1V0L	Hydrolase	1N5S	Oxidoreductase	1N6A	Transferase
1JK3	Hydrolase	1V3H	Hydrolase	1N62	Oxidoreductase	1NB9	Transferase
1JP4	Hydrolase	1W0P	Hydrolase	1N8K	Oxidoreductase	1NN5	Transferase
1JZ8	Hydrolase	1W3L	Hydrolase	1O7G	Oxidoreductase	1O7Q	Transferase
1KA1	Hydrolase	2NLR	Hydrolase	1ODM	Oxidoreductase	1OD6	Transferase
1KEI	Hydrolase	3MAN	Hydrolase	1PJ6	Oxidoreductase	1OE8	Transferase
1KIC	Hydrolase	6CEL	Hydrolase	1PZG	Oxidoreductase	1P0H	Transferase
1KWF	Hydrolase	1BD0	Isomerase	1QMG	Oxidoreductase	1P5Z	Transferase
1L8N	Hydrolase	1BKF	Isomerase	1RA2	Oxidoreductase	1Q0N	Transferase
1LJN	Hydrolase	1CCW	Isomerase	1RRM	Oxidoreductase	1Q36	Transferase
1LLF	Hydrolase	1IYH	Isomerase	1S2A	Oxidoreductase	1QD1	Transferase
1LO7	Hydrolase	1NYW	Isomerase	1UOG	Oxidoreductase	1QK3	Transferase
1ME4	Hydrolase	1OH0	Isomerase	1US0	Oxidoreductase	1QPC	Transferase
1MXG	Hydrolase	1OI6	Isomerase	1VLB	Oxidoreductase	1RDQ	Transferase
1N1T	Hydrolase	1P1J	Isomerase	3DFR	Oxidoreductase	1T46	Transferase
1N3Z	Hydrolase	1UDC	Isomerase	7ATJ	Oxidoreductase	1U4B	Transferase
1N9B	Hydrolase	1CSH	Lyase	1AJS	Transferase	1UU3	Transferase
1NF9	Hydrolase	1EEX	Lyase	1B4P	Transferase	1V2X	Transferase
1NSC	Hydrolase	1ELU	Lyase	1B8O	Transferase	1VHT	Transferase
1O6G	Hydrolase	1F74	Lyase	1BX4	Transferase	1VHW	Transferase
1O6I	Hydrolase	1KM6	Lyase	1CG6	Transferase	1V25	Ligase
1O7J	Hydrolase	1LXK	Lyase	1E19	Transferase	12AS	Ligase

1ODZ	Hydrolase	1M7Y	Lyase	1E2K	Transferase	1KQP	Ligase
1OGO	Hydrolase	1OC2	Lyase	1EJ0	Transferase	2A84	Ligase
1OJJ	Hydrolase	1OFL	Lyase	1EYN	Transferase	1GSA	Ligase
1OWE	Hydrolase	1R6D	Lyase	1FP2	Transferase	1BS1	Ligase
1PI5	Hydrolase	1R6W	Lyase	1G3M	Transferase	1GIM	Ligase
1Q0R	Hydrolase	1RWH	Lyase	1GHE	Transferase	1GPM	Ligase
1Q4U	Hydrolase	1SJW	Lyase	1GX5	Transferase	2QF7	Ligase
1Q74	Hydrolase	3STD	Lyase	1HNJ	Transferase	1XNY	Ligase
1Q92	Hydrolase	1AOE	Oxidoreductase	1HYV	Transferase	1A0I	Ligase
1QGI	Hydrolase	1CRU	Oxidoreductase	1I12	Transferase		
1QH5	Hydrolase	1D3G	Oxidoreductase	1I1N	Transferase		
1QHO	Hydrolase	1DF7	Oxidoreductase	1ID0	Transferase		

3.2.3 Predicting Enzyme Classes Methodology

To give a good prediction of the Enzyme Class, a score should be calculated for each triplet type indicating the tendency of this triplet to occur in the binding sites of a certain enzyme class. Six score tables are created (one for each enzyme class) containing the tendency of a triplet to occur in the binding sites of different enzyme classes. This tendency will be called the Enzyme Class Propensity (ECP). The ECP is calculated similarly to the triplet propensities in Equation 2-1, but this time, the comparison is done between the frequency of occurrence of a certain triplet in the binding sites of a certain class of enzyme and the rate of occurrence of that triplet in the binding sites of all enzyme classes in the training sets. This is shown in Equation 3-2.

$$\text{ClassProb}(\Phi, \alpha) = \frac{\text{ClassCount}(\Phi, \alpha)}{\sum_{i=1}^{455} \text{ClassCount}(\Phi, i)}$$

$$\text{EnzymeProb}(\Phi, \alpha) = \frac{\text{EnzymeCount}(\Phi, \alpha)}{\sum_{i=1}^{455} \text{EnzymeCount}(\Phi, i)}$$

$$\text{ECP}(\Phi, \alpha) = \log 2\left(\frac{\text{Class } \phi \text{ Prob}(\Phi, \alpha)}{\text{EnzymeProb}(\Phi, \alpha)} \right)$$

Where:
α denotes a triplet type;;
ClassProb(Φ, α) is the proportion of all triplets in Φ-class-binding-sites that are of type α;
EnzymeProb(Φ, α) is the proportion of all enzyme-binding-site triplets that are of type α;
ClassCount(Φ, α) is the count of occurrences of triplet type α in Φ-class ligand binding interfaces in the dataset;
EnzymeCount(Φ, α) is the count of occurrences of triplet type α in binding sites of all enzymes in the dataset.;
"i" spans the 455 triplet type;.
ECP(Φ, α) is the Enzyme Φ-Class Propensity score for a triplet Type α.

Equation 3-2: The Enzyme Class Propensity (ECP)

After the six ECP score tables have been generated, classifying the enzyme class of a certain enzyme is simply a matter of finding which score table gives the binding site of this enzyme the highest score. The binding site triplets are extracted and scored with all six ECP score tables. Each binding site receives six overall scores (one from each ECP table); each score being the average of the ECP scores of all the triplets in that binding site based on an ECP score table. The maximum of these six overall scores signals the class of the enzyme. For example, if the highest score (among the six overall scores) for a certain binding site is that generated from the hydrolase ECP score table, then that binding site is classified as a hydrolase.

3.2.4 Performance of STP in predicting Enzyme Classes

The enzyme classification system is tested on the structures making up its training set. The test is not performed according to the 90-10 testing scheme like the other STP and STPwater programs and that is because of the limited training set in use. The scarcity of isomerases, ligases, and lyases plays a key role in not adopting the 90-10 testing scheme, which would then mean creating an ECP scoretable for isomerases from merely 8 isomerase structures to test the 9th, which will include a large loss of the information held by the triplet propensity scores. Therefore the training set is tested fully with the generated ECP score tables.

STP showed a first level EC number identification success of 65%, correctly identifying 137 structures, and failing to identify 74 structures (Table 3-5). The success rate for each enzyme class can also be measured. Hydrolases are classified correctly in 66% of the test cases. That compares with 64% for the transferases, 59% for the oxidoreductases, 57% for the lyases, 82% for the ligases, and 78% for the isomerases (Figure 3-13). The high success rate for ligases and isomerases is possibly due to the low number of structures for these two classes in the training dataset (11 ligases and 9 isomerases) while the large classes (transferases, hydrolases, and oxidoreductases) show success rates closer to the entire subset success rate. Further scrutiny shows that 44% of the wrongly predicted oxidoreductases were classified as transferases, 44% of the wrongly predicted hydrolases were classified as ligases, and 33% of the wrongly predicted transferases were classified as hydrolases (Figure 3-12). This indicates a similarity between the profiles of these enzymes and will be discussed further in Section 3.2.5.

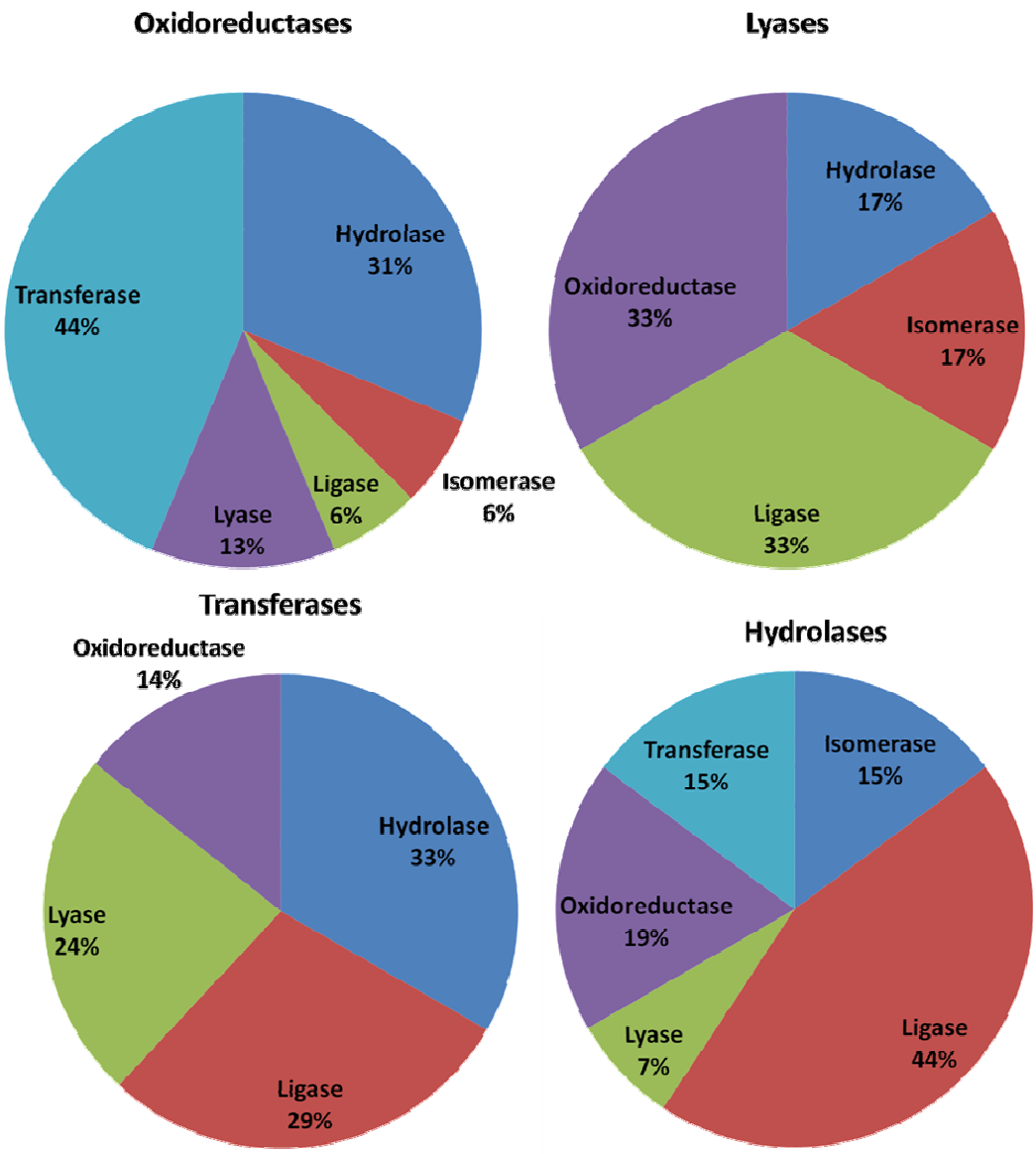

Figure 3-12: The distribution of the false predictions of the four largest enzyme classes. Ligases and isomerases had two false predictions each and therefore the statistics in these cases are unreliable. Results show a high tendency to mistake hydrolases for ligases, oxidoreductases for transferases and hydrolases, transferases for hydrolases.

Table 3-5: The 74/211 Structures with failed predictions of their EC Classes by STP.

PDB ID	Enzyme Class	STP Prediction	PDB ID	Enzyme Class	STP Prediction
1BUP	Hydrolase	Ligase	1CRU	Oxidoreductase	Lyase
1BXO	Hydrolase	Isomerase	1FRB	Oxidoreductase	Hydrolase
1DL2	Hydrolase	Lyase	1G8K	Oxidoreductase	Transferase
1E4M	Hydrolase	Ligase	1HFU	Oxidoreductase	Transferase
1F8E	Hydrolase	Ligase	1IW0	Oxidoreductase	Lyase
1GG6	Hydrolase	Ligase	1KB0	Oxidoreductase	Hydrolase
1GKL	Hydrolase	Transferase	1LC3	Oxidoreductase	Isomerase
1GNX	Hydrolase	Ligase	1LQT	Oxidoreductase	Transferase
1J1M	Hydrolase	Ligase	1N62	Oxidoreductase	Transferase
1JK3	Hydrolase	Ligase	1N8K	Oxidoreductase	Transferase
1LLF	Hydrolase	Oxidoreductase	1O7G	Oxidoreductase	Hydrolase
1LO7	Hydrolase	Ligase	1PZG	Oxidoreductase	Transferase
1ME4	Hydrolase	Ligase	1QMG	Oxidoreductase	Ligase
1N1T	Hydrolase	Lyase	1RRM	Oxidoreductase	Transferase
1N9B	Hydrolase	Transferase	1US0	Oxidoreductase	Hydrolase
1NF9	Hydrolase	Oxidoreductase	7ATJ	Oxidoreductase	Hydrolase
1O6G	Hydrolase	Isomerase	1AJS	Transferase	Lyase
1O7J	Hydrolase	Ligase	1E2K	Transferase	Oxidoreductase
1PI5	Hydrolase	Ligase	1FP2	Transferase	Hydrolase
1Q0R	Hydrolase	Oxidoreductase	1HYV	Transferase	Ligase
1Q4U	Hydrolase	Transferase	1J54	Transferase	Hydrolase
1QGI	Hydrolase	Transferase	1JG1	Transferase	Ligase
1SU2	Hydrolase	Oxidoreductase	1JVP	Transferase	Oxidoreductase
1TBF	Hydrolase	Oxidoreductase	1JX4	Transferase	Ligase
1UTP	Hydrolase	Ligase	1KJQ	Transferase	Lyase
1UWC	Hydrolase	Isomerase	1LZJ	Transferase	Hydrolase
1W0P	Hydrolase	Isomerase	1M15	Transferase	Lyase
1NYW	Isomerase	Hydrolase	1M4I	Transferase	Hydrolase
1OH0	Isomerase	Oxidoreductase	1N6A	Transferase	Hydrolase
1A0I	Ligase	Lyase	1NB9	Transferase	Hydrolase
1XNY	Ligase	Oxidoreductase	1O7Q	Transferase	Hydrolase
1CSH	Lyase	Ligase	1OE8	Transferase	Ligase
1F74	Lyase	Isomerase	1Q0N	Transferase	Ligase
1KM6	Lyase	Ligase	1Q36	Transferase	Lyase
1RWH	Lyase	Hydrolase	1QD1	Transferase	Lyase
1SJW	Lyase	Oxidoreductase	1QK3	Transferase	Ligase
3STD	Lyase	Oxidoreductase	1T46	Transferase	Oxidoreductase

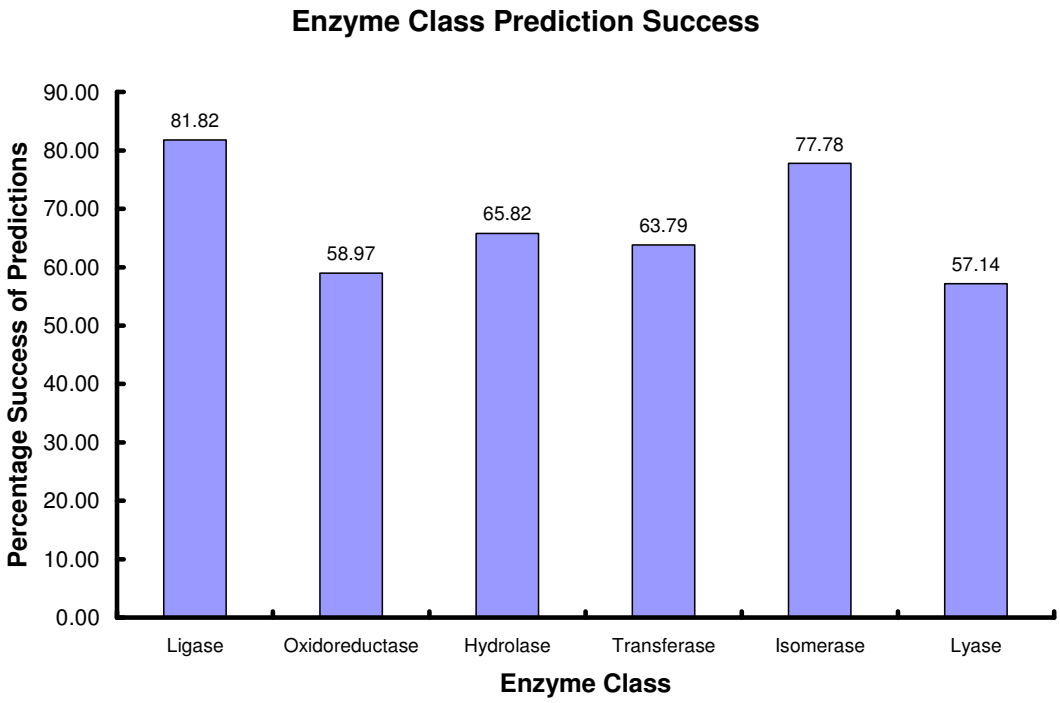

Figure 3-13: The performance of STP in predicting the Enzyme Classes of the 6 Enzyme Classes.

This performance is compared to that of the method designed by Dobson and Doig (2005) [186]. This method relies on a support vector machine (SVM) trained on a non-redundant dataset to classify enzymes by their first-level E.C. Number. The SVM takes as input several attributes related to the geometry and secondary structure of the tested enzyme (surface area, volume, Stride [187] assignments of secondary structure, fractal dimension [188], and number of residues). This method succeeds in predicting the EC number of 35% of the tested enzymes. In contrast, STP succeeds in predicting the EC numbers in 65% of the tested cases. This advantage posed by STP is possibly due to the nature of the STP algorithm which relies on the triangular composition of the pocket under study. These triangular patterns implicitly define

physiochemical attributes of that pocket and lead to a fast and reliable classification of the enzyme.

3.2.5 One versus One Class Predictions

With the existence of six enzyme classes, the problem of one versus one class predictions is defined as the classification of a certain structure into any of two possible classes (rather than a possible six). This is applicable when the enzyme in question is suspected of being any of two classes and is thus tested accordingly. Dobson and Doig (2005) [186] refers to this problem as the 15 one class versus one class problems (as there are 15 combinations of 2 classes out of 6) and they report their success rates (Table 3-6). The same experiment is conducted with use of the ECP score tables where each of the 15 one versus one problems is tested for the accuracy. For example, testing Hydrolases versus Lyases means all Hydrolases and Lyases are scored with the Hydrolase and Lyase ECP score tables. Each structure is then classified according to the table that gives it the higher score. The results of these 15 one versus one problems are listed and compared to those of Dobson and Doig (2005) [186]. STP-ECP scores perform better than the Dobson and Doig [186] method in each of the 15 one versus one class predictions, except in the case of Oxidoreductase versus Hydrolase class prediction, where the success rate of Oxidoreductase is 76.9% with STP-ECP score tables and 79.7% with Dobson and Doig [186]. However, the total success rate for the Oxidoreductase versus Hydrolase is higher for STP-ECP score tables (77.1%) than for the other method (67.4%).

Table 3-6: Predicting Enzyme Classes (first-level EC numbers) divided into 15 one class versus one class problems. The performance of the STP-ECP (Enzyme Class Propensity) score tables is compared to that of the method by Dobson and Doig (2005) [186]. In bold is the only statistic where Dobson and Doig's method [186] outperforms STP-ECP. In a problem of A vs B, accuracy of A is the percentage of A instances predicted correctly, and similarly for B. Total accuracy is the percentage of all A or B instances predicted correctly.

Problem		Accuracy (%)					
		Dobson and Doig [186]			STP-ECP score tables		
A	B	A	B	Total	A	B	Total
Oxidoreductase	Transferase	68.4	64.8	66.2	79.5	87.9	84.5
Oxidoreductase	Hydrolase	79.7	61.3	67.4	**76.9**	77.2	77.1
Oxidoreductase	Lyase	75.9	75.0	75.5	94.7	85.7	92.5
Oxidoreductase	Isomerase	73.4	74.5	73.8	87.2	88.9	87.5
Oxidoreductase	Ligase	81.0	75.0	79.8	84.6	90.9	86.0
Transferase	Hydrolase	58.6	58.8	58.7	82.8	77.2	79.6
Transferase	Lyase	35.9	75.0	48.4	81.0	85.7	81.9
Transferase	Isomerase	53.9	66.7	57.5	84.5	100	86.6
Transferase	Ligase	59.4	75.0	61.4	84.5	90.9	85.5
Hydrolase	Lyase	46.2	78.3	55.0	78.5	82.9	80.6
Hydrolase	Isomerase	58.8	68.6	61.1	78.5	88.9	79.5
Hydrolase	Ligase	49.4	70.0	51.7	75.9	100	78.9
Lyase	Isomerase	50.0	68.6	58.6	92.9	100	95.7
Lyase	Ligase	50.0	60.0	52.4	85.7	90.9	88.0
Isomerase	Ligase	62.7	70.0	64.8	100	100	100

The lowest total accuracies scored by STP-ECP (Table 3-5) were recorded (by increasing accuracy) for the oxidoreductase : hydrolase, hydrolase : ligase, hydrolase: isomerase, transferase: hydrolase, and hydrolase:lyase sub-problems. Interestingly, all these problems included hydrolases as a candidate classification, indicating that hydrolases are more prone to being confused with other classes. Combined with the observations in Figure 3-12 (Section 3.2.4), we see that 33% of the wrongly predicted transferases were classified as hydrolase. The same goes for 31% of the wrongly predicted oxidoreductases and 17% of the wrongly predicted lyases. We conclude that hydrolase active sites are less distinct from the other enzyme classes according to the STP-ECP score tables. Nevertheless, the accuracy

scores are encouragingly high, with the lowest total accuracy score (for the oxidoreductase:hydrolase subproblem) at a value of 77.2%.

3.3 Using STP Propensities to Rank Docking Orientations

Protein-protein docking programs generate a large and varied list of possible docking orientations between two binding partners. The energy/shape functions in these programs are sometimes insufficient to discriminate between orientations, leading to the true docking pose being absent from the best hits. This section demonstrates the use of STP to re-rank such docking orientations in an attempt to get a better prediction of how two complement proteins, C5 and C7, interact during mammalian immunological response.

3.3.1 The HyHEL-10 Fab-lysozyme Complex

Table 3-7: Performance of Hex in the recognition of known complexes. A 6D-search over 5.4×10^8 alternative test orientations is carried out at increasing fourier transformation expansion orders, N. The lowest energy orientation within 3 Å (calculated as that of the C_α deviations) from the experimental ligand location is taken and its rank is noted. [72].

PDB	N=16		N=20		N=25	
	Hex Rank	RMS (Å)	Hex Rank	RMS (Å)	Hex Rank	RMS (Å)
2DHB	2	0.00	2	0.00	1	1.55
2CCY	1	0.04	1	0.04	1	1.59
1CSE	37	0.73	1	0.08	1	0.92
2SNI	15	0.58	1	0.42	1	0.42
2KAI	17	0.41	3	0.69	7	0.81
2PTC	132	0.52	2	0.48	1	0.48
1CGI	1	0.38	1	0.38	1	0.38
1CHO	1	0.45	1	0.55	1	0.55
1BGS	1	0.82	1	0.82	1	0.88
1GGI	1	2.47	1	0.90	1	0.90
1TET	5	1.48	1	1.16	1	1.09
1FPT	102	1.04	1	0.42	1	0.42
2IGF	3	0.71	1	0.77	1	0.77
1JEL	4867	0.81	1060	0.81	2	0.81
1BQL	524	1.85	12	0.96	1	0.39
3HFL	318	1.01	5	1.00	1	1.00
3HFM	7	2.19	27	1.09	3	1.03
1VFB	8344	1.49	216	0.20	9	0.20
1MLC	1401	0.00	116	0.00	187	0.84

1MEL	9898	1.03	27	1.03	3	1.03
1JHL	385	0.62	8	0.38	1	1.08
1FBI	14	1.09	1	1.09	1	0.38
1NCA	68	1.53	1	0.32	1	0.32
1NMB	160	2.43	1630	1.39	1009	1.39
1NSN	19992	1.11	716	0.75	1130	2.29
1IAI	1381	1.48	111	0.37	20	1.39
1DVF	11145	0.00	88	1.38	49	0.44
1KB5	140	0.34	1	0.34	78	1.38
1IGC	1328	1.74	269	0.81	1	0.34

Hex 5.1 [72] is the main docking engine for this experiment. Hex uses spherical polar Fourier correlations to define the protein surfaces and tries over 540 million docking orientations for each pair of docking partners. Those orientations are scored based on shape complementarity, electrostatic interactions, and molecular mechanics. Hex performs remarkably well in most cases of docking globular domains (Table 3-7). However, as reported by Ritchie and Kemp [72], Hex sometimes fails to identify the right orientation. Frequently, the correct orientation is examined by Hex during the global search, but is not ranked highly according to the shape complementarity algorithm. This shortcoming can been corrected using STP to rerank the orientations [133].

Table 3-8 A snapshot of the top 5 HyHEL-10 Fab-Lysozyme docking orientations after being ranked by STP shows that STP finds the best orientations unambiguously [133]. Interface triplets on both receptor and partner are identified, and the average propensity is calculated as a score for the orientation. The orientation with the highest average interface propensity is ranked first.

Dock ID	Hex Rank	STP Average Propensity (Rank)	RMS from Crystal Structure (Å)
Dock0058	58	0.20 (1)	1.27
Dock0012	12	0.16 (2)	3.47
Dock0118	118	0.15 (3)	19.94
Dock0114	114	0.14 (4)	14.36

We have previously demonstrated a similar experiment where ranking docking orientations were performed on the HyHEL-10 Fab-lysozyme complex [133] . In that case, the three-dimensional structure of the complex between the antibody and lysozyme was known. The key results, listed in Table 3-8, show that the true docking orientations, which were ranked 58th and 12th by Hex, were ranked first and second by STP.

3.3.2 The C5/C7 interaction of the Human Immune Complement System

We have used the ranking procedure to carry out a detailed study of the interaction between the C5 and C7 subunits of the immune complement system. A complement-mediated response to infection is vital to recognize, and kill, pathogens and toxic entities. Inappropriate complement activity, whether activation at inappropriate sites or at excessive/inadequate levels, leads to numerous inflammatory disorders and/or tissue damage [122, 189]. The activation of complement pathways and thus, the attack on pathogens, is a result of a cascade of intermolecular binding, enzymatic cleavages and protein complex assembly [121]. Although many of the binding partners and their roles have been identified [190, 191], there is still very little understanding of the molecular level interactions and how those binding partners behave in order to create the necessary complexes needed for the complement-mediated immune response. Specifically, although the C5-C6-C7-C8-C9 proteins which form the lipid bilayer penetrating attack complex (Figure 3-14) are known, the atomic-level details of interactions between them are still unknown [121].

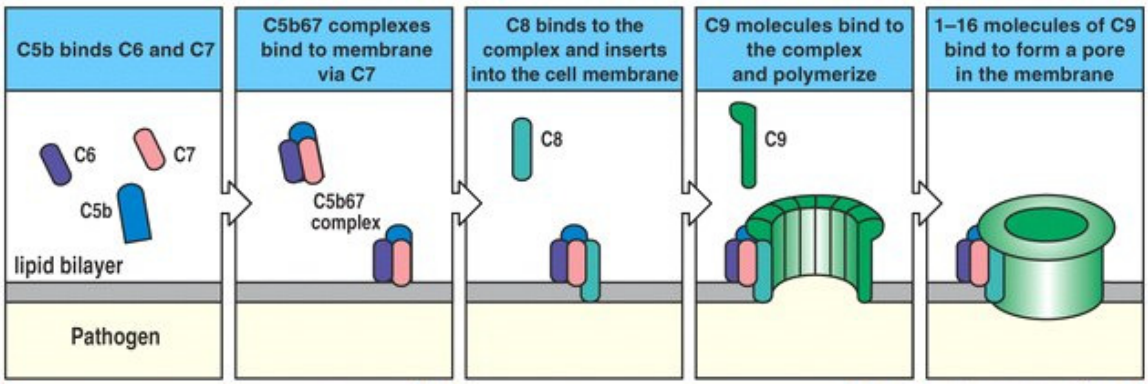

Figure 3-14: The Membrane attack complex formation, picture taken from [192], Fig 2-35.

The membrane attack complex (MAC) is assembled by C5 binding to C6, and then to C7 and C8 [121]. Then a membrane pore is formed by the assembly of several copies of C9 to the initial complex. The C-terminal domain of C5, the C345C domain, has been shown to interact with both C6 and C7, through their C-terminal factor I-like modules (FIMs [193]). Bramham et al. have determined the solution structures of the C5-C345C domain [191] (Figure 3-15) and C7-FIMs [194] (Figure 3-16) in isolation but not yet in complex. They are interested in finding how these two proteins interact since characterizing the intermolecular details of the interaction between C5 and C7 should provide insight into how the MAC begins to assemble. This valuable information should also shed some light on possible ways to fight the complications that arise from inappropriate complement activity in terms of how to up/down-regulate or inhibit the formation of the MAC. Gasque et al [122] provide evidence of the involvement of complement in the initiation/exacerbation of central nervous system inflammation and tissue injury, and suggests that successful inhibition of the complement system in these cases is of therapeutic value.

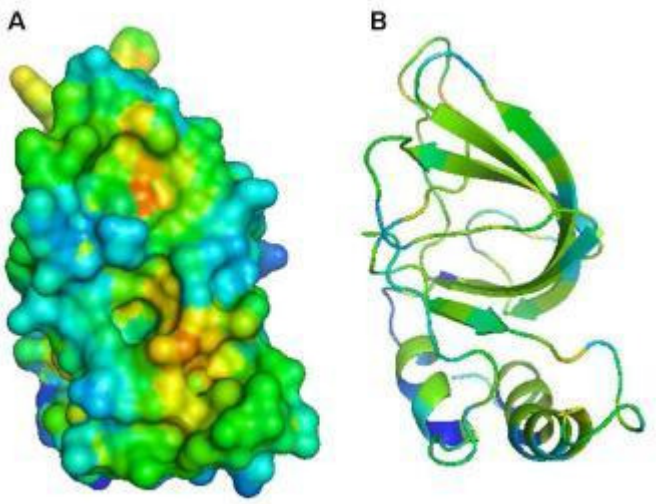

Figure 3-15: The structure of the C5-C345C domain in surface (A) and cartoon (B) views, and colored by STP

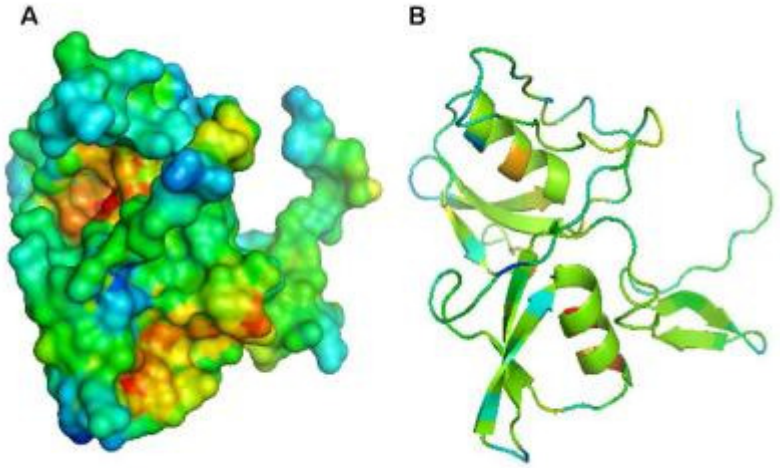

Figure 3-16: The structure of the C7-FIMs in surface (A) and cartoon (B) views, and colored by STP

Three docking experiments were designed to explore the possible ways of interaction between C5-C345C and C7-FIMs, and the different parameters are shown in Figure 3-17. Electrostatic interactions were always used in all the dockings. However, the post processing step was varied within the docking simulations between “Bumps and

Volumes", "MM Minimization", and "MM Energy". The first option enables a bumps counter, in which the number of steric clashes between non-bonded pairs of heavy atoms in each solution is calculated and used in scoring the docking orientations. The second option utilizes a single (rigid body) molecular mechanics energy for scoring those orientations. The third option applies a Newton-like energy minimization to each docking solution. These energies are calculated using Lennard-Jones and hydrogen bond potentials as defined in the OPLS forcefield in addition to electrostatic interactions (Hex 5.1 user manual).

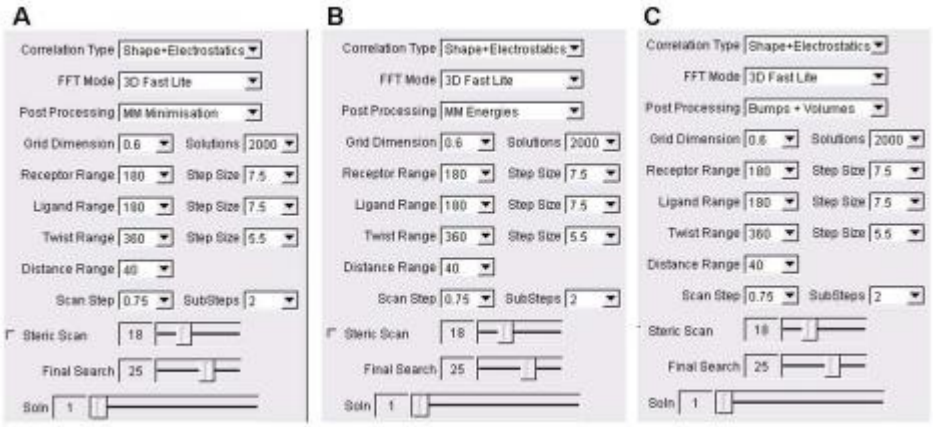

Figure 3-17: The Hex parameters used to dock C5-C345C onto C7-FIMs in the three experiments denoted as A, B, and C

Hex was set to return the top 500 solution clusters in each experiment, where similar docking orientations (based on calculating RMS distances between the different poses of the ligand domain in different orientations) are automatically grouped in clusters. These 500 solutions were subsequently scored via STP. The binding site triplets and atoms were extrapolated, and each solution orientation was given two scores: a triplet score equal to the average of STP propensity scores of all triplets in the binding site, and an atom score equal to the average of STP propensity scores of

all atoms in the binding site. Docking orientations that were ranked in both the top 100 results by Hex and the top 100 by STP triplet scores or STP atom scores were classified as "spotlight" orientations to be subjected for further analysis.

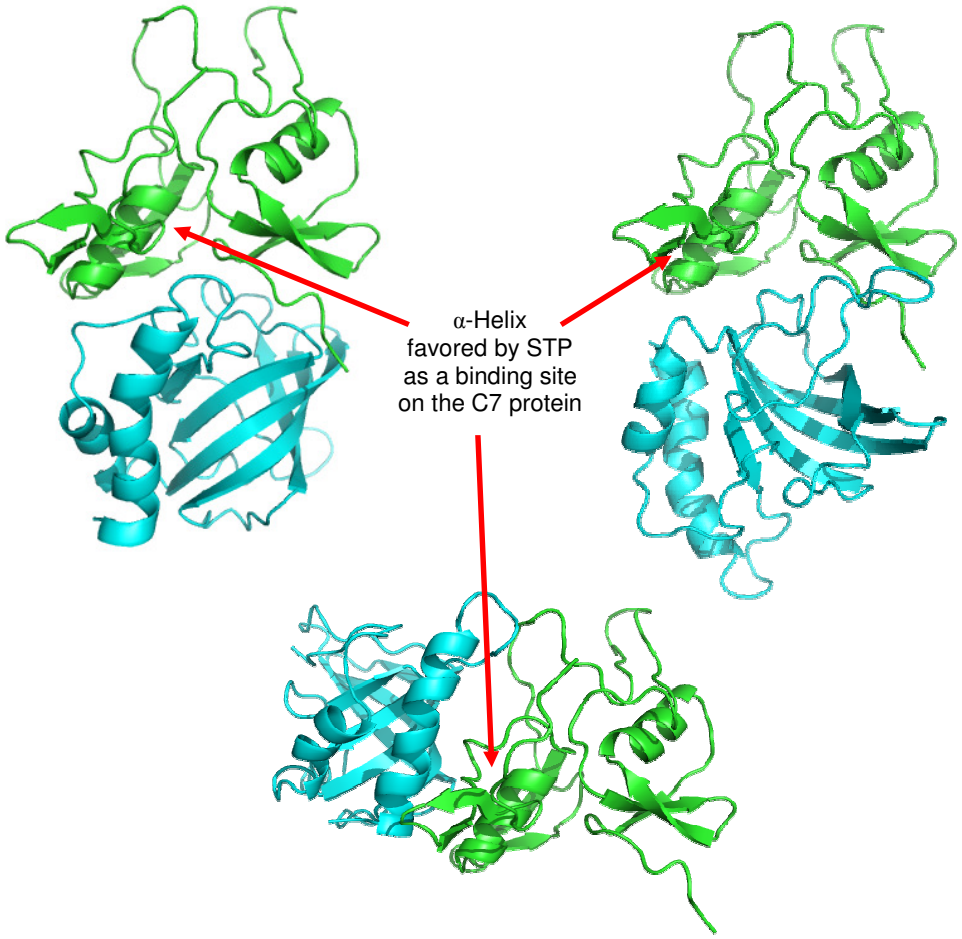

Figure 3-18: Representative orientations of different docking clusters of the C7-FIMs (green) and C5-C345C (blue) proteins showing the top hits according to the STP/Hex combined score involving the Val 54 to Gln 63 α-helix that is favored by STP as a binding site. In most orientations, the interaction involves both modules of the C7-FIM pair (top 2 orientations), however some orientations were on the other side of the α-helix which also receives a high STP score (bottom orientation).

In this case, an assessment of the success of the STP predictions of the binding sites is not an easy task due to the absence of a three-dimensional structure of the C7/C5

complex. Hex produced a large number of orientations for the C5/C7 complex, but conjugating the Hex scores with STP scores led to the favoring of orientations taking place on either side of the α helix extending from Val 54 to Gln 63 in the C7-FIMs (Figure 3-18).

An NMR titration experiment was designed to determine which residues of the C7-FIMS were perturbed upon binding to C5, by examination of the resonances from the backbone amide groups in series of HSQC spectra. Although in this titration the NMR experiment does not show the location of the C5 protein in the complex, it indicates the likely location of the interface. This experiment showed that the strongest effect occurs on and around the α-helix extending from Val 54 to Gln 63 in the C7-FIMS; the same helix predicted by STP to be a binding site (Figure 3-19). This result is also backed up by the fact that the C7 fragment in this study constitutes a pair of factor I-like modules in a compact pseudosymmetric arrangement [194]. It has been suggested that this pair of domains might undergo a structural alteration upon binding, opening up to a more elongated conformation to accommodate a binding partner during the assembly of the MAC [194].

STP predictions concur with the titration experiment. The side chains of all residues colored in orange, red, and magenta in the titration experiment (Figure 3-19) have been colored orange and red by STP. The results of the titration experiment show the importance of using STP scores in ranking docking orientations. With the help of

STP and Hex, we are now one step closer in understanding the complex nature of the MAC self-assembly process.

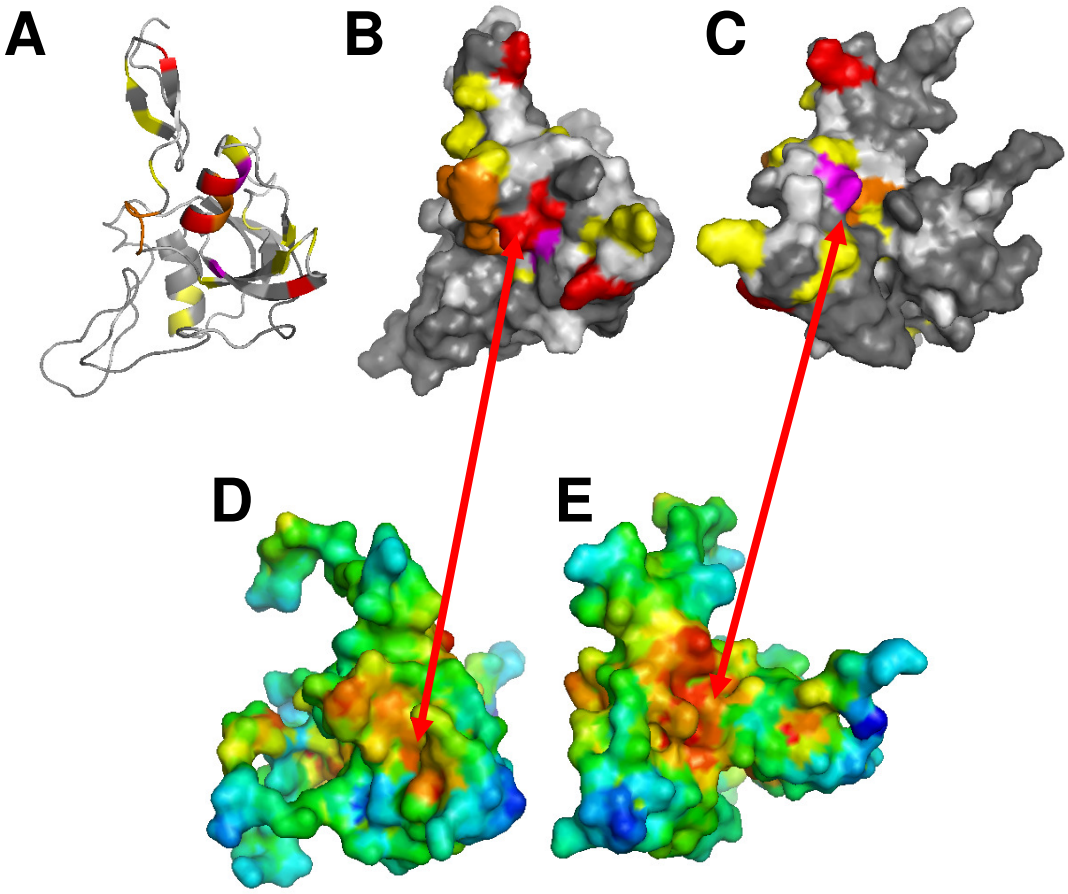

Figure 3-19: The sites on C7-FIM likely to interface with C5-C345C. A, B, and C show the results of a titration experiment of C5-C345C into C7-FIM where the perturbation of the backbone amides upon binding to C5 was measured. Color spectrum is grey, white, yellow, orange, magenta, and red (in increasing order of perturbation). Residues affected most are expected to be either in the binding interface, or linked to other residues in the interface via secondary structure, or to have been displaced by interdomain rearrangement. The sites shown in B and C are opposite sides of an α-helix. The STP colored variants of B and C are shown in D and E respectively. STP successfully identifies the same sites as the NMR titration.

3.4 Identifying Protein-Protein Binding sites in large multi-component complexes: the Cks1-Dependent recognition of p27 by the SCF-SKP2 Ubiquitin Ligase

Molecular functions are a result of interaction involving any number of proteins. Many assemblies include more than 2 subunits. Examples include the VHL/elonginC/elonginB complex (PDB 1vcb), where the tumor suppressor protein VHL inhibits the formation of the elonginA/elonginB/elonginC complex, a main reason behind the von Hippel–Lindau disease and the majority of kidney cancers [195]. Other large complexes are the 10-subunit RNA polymerase II elongation complex (PDB 1I6H) [196], the 7-subunit arp2/3 complex (PDB 1K8K) responsible for formation of Y-branch actin filaments and the motion in eukaryotic motile cells [197], and the 3-subunit MHC-II/TCR/SpeA complex [198] responsible for the activation of the immune system [81].

The SCF-SKP2 Ubiquitin E3 Ligase comprises five subunits: Cul1 domains 1 and 2, SKP1, SKP2, and CKS1 (Figure 3-22). Successful Inhibition or regulation of such complexes depends on understanding the binding details of these complexes. Such details may include function, chronological order of binding, and strength of binding. We demonstrate the ability of STP to quickly highlight the binding strength between pairs of domains within large complexes, giving a clearer idea of how to regulate them.

When the cell undergoes a transition from the G1 to the S phase, the commitment of the cell to the S phase is dependant on the activation of any the cyclin dependent kinases Cdk2/E or Cdk2/A [199]. Cdk2/E and Cdk2/A are responsible for down regulating p27 and that is a necessary step for the G1/S cell cycle transition [200]. Studies have shown that p27 becomes phosphorylated upon binding to Cdk2/E and that the down regulation of p27 takes place via the recognition of phosphorylated p27 in the p27/Cdk2/E complex by the $SCF^{Skp2/Cks1}$ E3 Ligase which would then lead to the ubiquitination and degradation of p27 at the G1/S transition [201, 202]. This interaction is of great importance, and that is due to the fact that p27 has been linked with several aspects of cancer. DNA replication and formation of tumors was prevented in certain cancer cells on nude mice by the over expression of p27 [203]. Moreover, p27 loss is associated in many human cancers like breast, prostate, colon, gastric, lung, and esophageal cancers [204]. Abnormal increase in the degradation of p27 could be the reason behind the loss of p27 in cancer cells [205]. Understanding more about the p27 regulation process is thus of vital therapeutic value.

The interaction between $SCF^{Skp2/Cks1}$ and p27 has been identified and the structure of the Cul1-Skp1-Skp2-Cks1-p27 complex has been solved [206, 207]. This complex has been reported to have an increased expression in cancer cells (which itself leads to low p27 levels) [208, 209]. The base SCF complex is formed by the binding of Skp1, Cul1, Rbx1, and the Fbox protein that binds to the substrate [210]. In this case, the Fbox protein is Skp2. The binding of Cul1, Skp1, and Skp2 is shown in Figure 3-20.

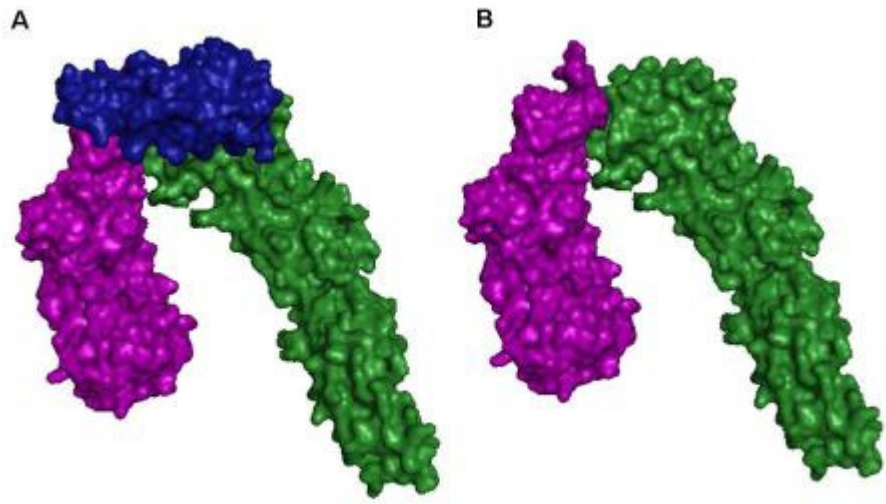

Figure 3-20: The structure of the SKP1 (BLUE) - SKP2 (magenta) - CUL1 (green) complex

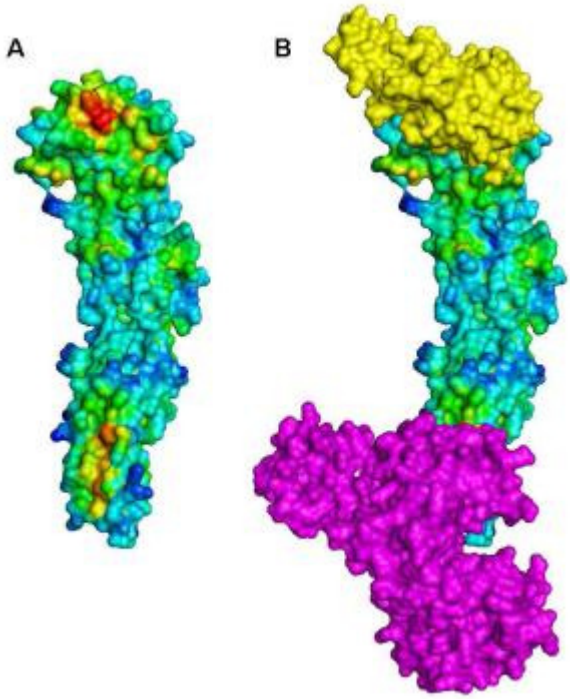

Figure 3-21: The coloring of Cul1 by STP and its respective binding partners. Skp1 is shown in Yellow and the second domain of Cul1 in Magenta

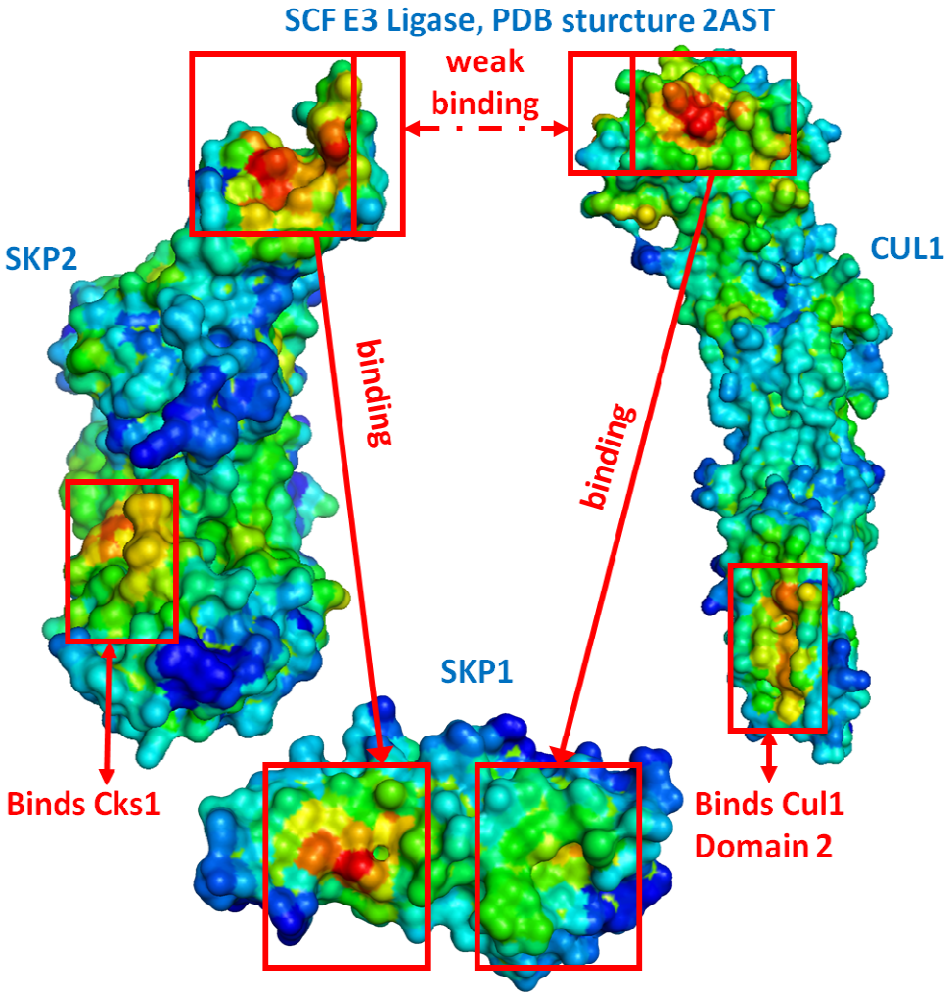

Figure 3-22: The SCF complex (PDB 2AST) chains colored by STP and the binding sites being marked. The coloring shows low binding affinity between SKP2 and CUL1, suggesting that SKP1 attaches the complex together, and that has been shown in the literature [207]

At first glance, Cul1 and Skp2 seem to be interacting. Upon coloring the surface of those subunits by STP, Cul1 shows 2 distinct binding sites; the first binds to Skp1 and the second binds the second subunit of Cul1 which in turn binds rbx1 (Figure 3-21). The surface of Cul1 that binds to Skp2 does not show a strong binding influence. In turn, the surface on Skp2 that binds to Cul1 is also of poor STP scoring.

Skp1 and Skp2 both show strong binding surfaces that bind to each other. Those details are shown in Figure 3-22. This suggests that Cul1 and Skp2 have poor binding affinities towards each other, but are clipped together by Skp1. This STP prediction is correct since Cul1 is reported by [207] to be the rigid scaffold which binds Skp1, which in turn binds to Skp2. Had the structure of the SCF complex been hard to crystallize as a single unit, STP would have helped reach this conclusion just from examining the individual structures of Cul1, Skp1, and Skp2.

STP has already been proven useful in locating binding sites. In this chapter, this functionality was taken one step further, with the location of several binding sites on each subunit of the SCF E3 Ligase and the comparison of the STP propensities of the triplets in these binding sites to reveal the strength of the binding at each interface. This shows that STP is useful not only in locating binding sites, but also in comparing them. For this reason, a special version of STP has been designed to color surfaces of several proteins on the same coloring scale (rather than each protein getting its own scale of 0-100). This facilitates the comparison of binding sites across different structures to make use of the STP functionality described in this section.

3.5 Predicting Allosteric Binding Sites

Allosteric binding sites form when a ligand binds to a protein, causing a structural change that leads to modified affinities at other (remote) ligand binding sites. Allosteric regulation controls many important cellular processes, including signal transduction, transcription, and metabolism [211]. Allosteric sites are difficult to locate both experimentally or computationally where molecular dynamics simulations need to be used. STP is helpful in this matter and can be used in two methods. First, the coloring of the protein surface often generates multiple highly scored patches. These patches can be investigated for being secondary or allosteric binding sites (Section 3.4). Alternatively, upon the existence of several conformational states identified through X-ray crystallography or dynamics, these states can be colored with STP and investigated for the formation of highly colored patches.

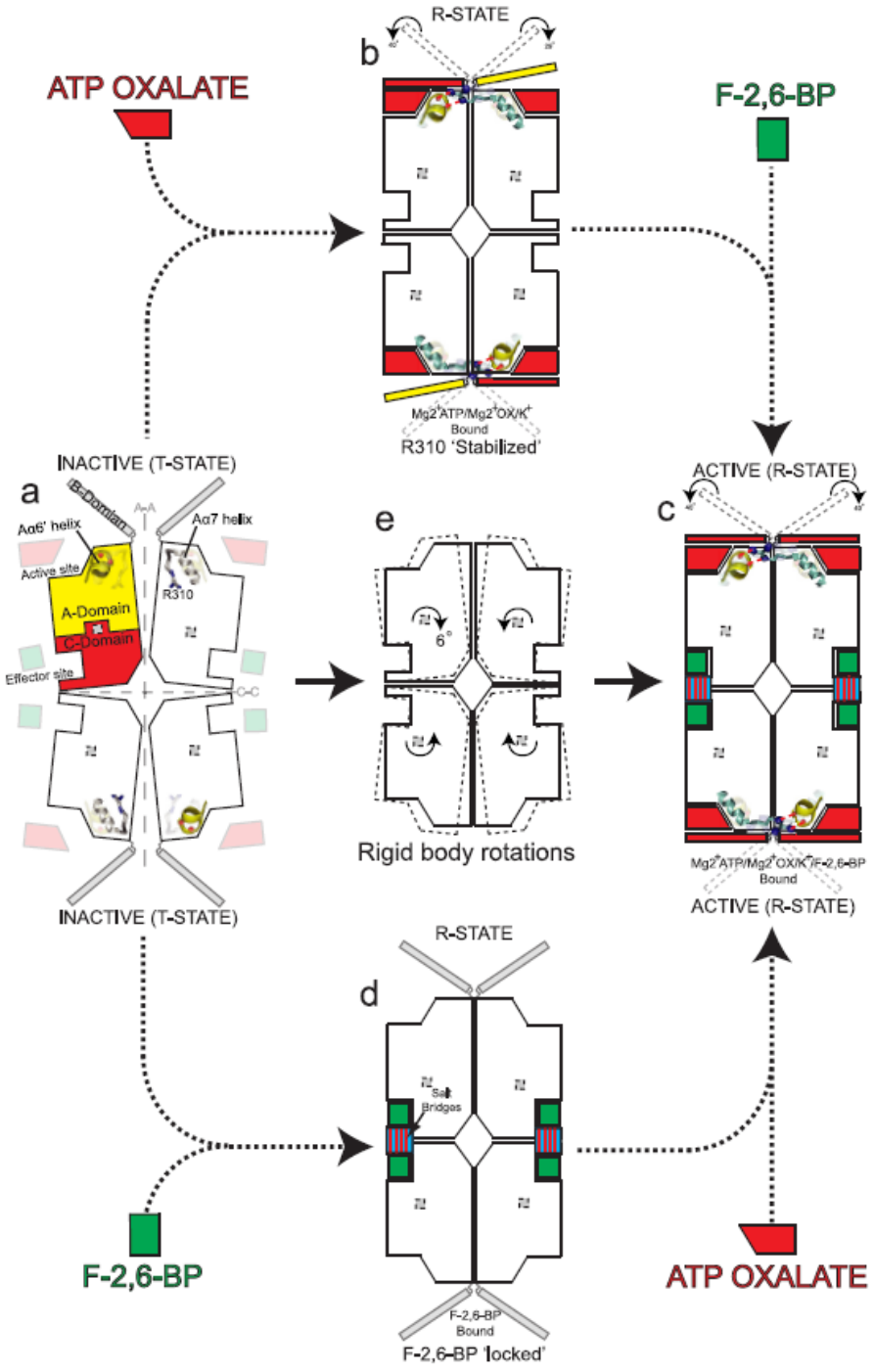

Figure 3-23: The transition of Lm-PyK between the inactive T state and the active R state, with the help of the allosteric ligand F-2,6-BP to stabilize the structure in the R-state. Picture taken from Morgan et al., (2010) [86].

We present an example, Pyruvate Kinase (PyK), an enzyme responsible for catalyzing the final reaction of glycolysis, where phosphoenolpyruvate (PEP) and ADP are converted into pyruvate and ATP, respectively [212]. PyK is a homotetramer, each monomer in the range of 50-60 kDa depending on species. The human embryonic and tumour (M2), erythrocyte (R), and liver (L) isoforms of PyK are allosterically activated by fructose 1,6-bisphosphate (F-1,6-BP) [86]. This enzyme for leishmania has been studied extensively and undergoes a transition between an inactive T-state and an active R-state (Figure 3-23), and the binding of fructose 2,6-bisphosphate (F-2,6-BP) in the effector sites of PyK leads to stabilizing the enzyme in the active R-state. PyK has been extensively studied and crystallized and studied in house by Hugh Morgan [86, 212, 213].

The comparison between the STP colored versions of the R-State and the T-State of Lm-PyK is shown in (Figure 3-24). The increase in color intensity of the binding sites, as well as the formation of new red patches on the R-State concurs with the biological studies showing that the R-state is the active state of the enzyme. The colored surface of the R-state is also compared with the biological data gathered through crystal trials [86, 212, 213] about the location of the effector and active sites of the enzymes. This comparison also demonstrates STP's success in picking out the allosteric sites of PyK (Figure 3-25).

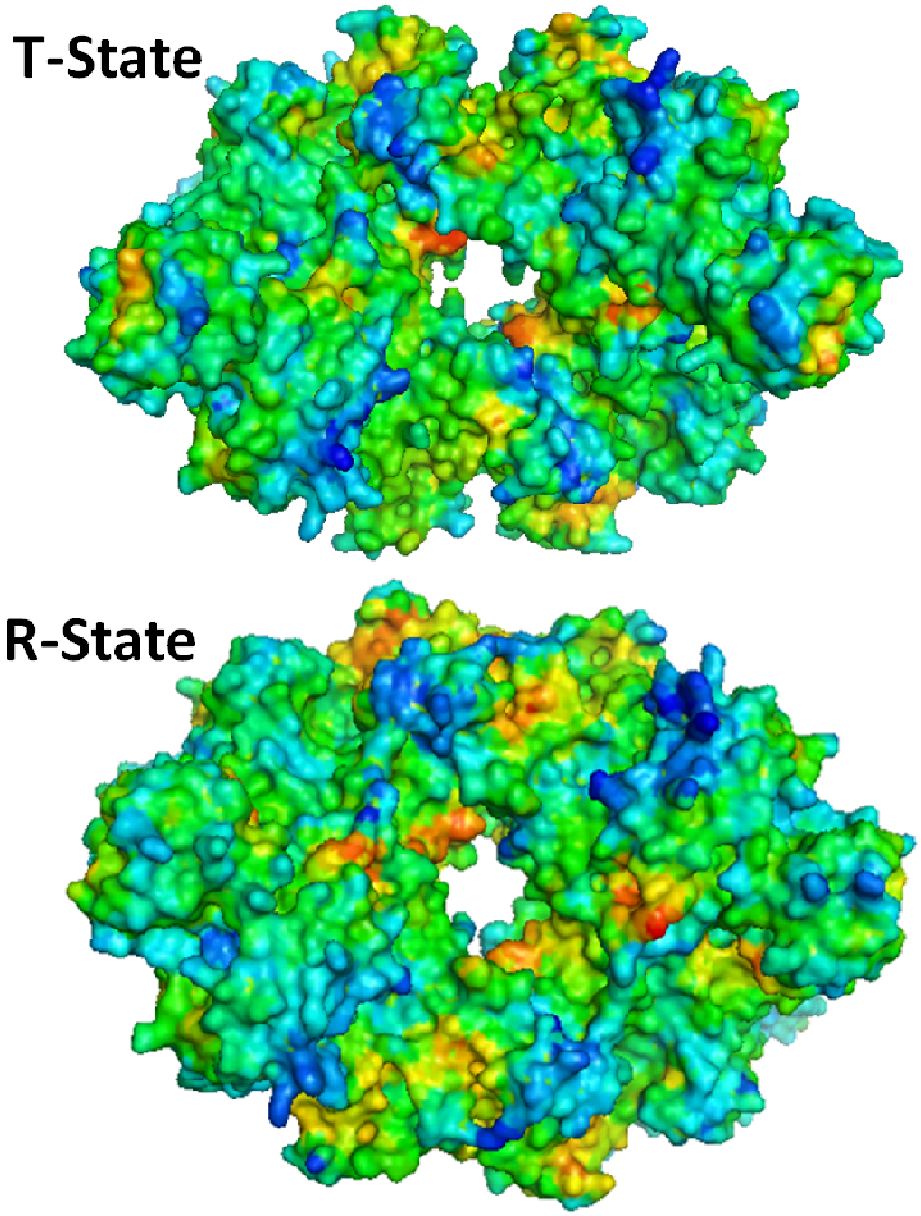

Figure 3-24: Comparison between the STP colored versions of the R-State and the T-State of PyK. The increase in color intensity of the binding sites, as well as the formation of new red patches on the R-State concurs with the biological studies showing that the R-state is the active state of the enzyme [86, 212, 213].

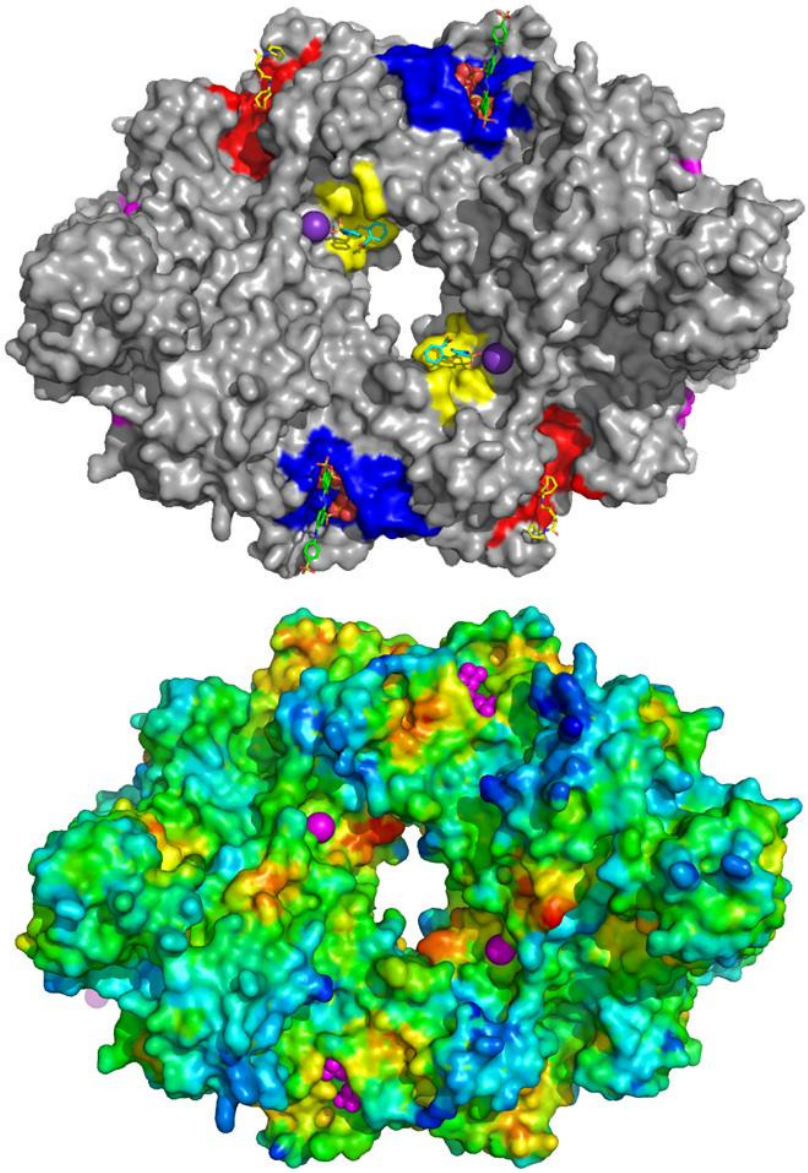

Figure 3-25: The allosteric sites on Pyruvate Kinase (structure solved by Hugh Morgan in the Walkinshaw Group). Top Figure shows the allosteric sites highlighted in blue, yellow, and red. The bottom picture shows the STP-colored surface, where the allosteric sites have been predicted.

4 Spatial and Chemical Features of Protein Surfaces

This chapter studies the spatial and chemical characteristics of protein surfaces from an STP perspective. Spatial features of protein surfaces in protein-ligand, protein-protein, and protein-peptide interaction datasets are studied: the surface triplets are generated and certain geometrical characteristics are calculated. Inter-triangular distances (distances between the centroids of triangles) are calculated for whole protein surfaces and binding sites. The distributions of these distances are compared to study the compactness of protein surfaces in binding sites. The shape of the triplets is also studied, through the areas of the triplets, and the lengths of their edges. We use these characteristics to draw conclusions about the concavity of binding sites and to compare the binding sites of the three interaction datasets.

The second part of this chapter studies the chemical features of triplets. Triplets are chemically classified based on their constituting atoms. The role of each chemical category of triplets in each of the interaction datasets in this work is studied and conclusions are drawn as to the importance of each of them in the three interaction types (protein-ligand, protein-protein, and protein peptide). The tendency of each of these chemical triplet categories to interact with ligand atom types is also studied and quantified (interaction preference values). These statistics are used to calculate a statistical free energy of interaction between triplets and ligand atoms. The interaction preference and statistical free energies give rise to interesting conclusions and provide quantitative proofs of many existing postulates of interaction (for eg, the preference of a hydrophobic atom to exist in a hydrophobic region).

4.1 Inter-triangular Distances

4.1.1 Protein-Ligand Interaction Dataset

This section details the studies aimed at assessing the distinct spatial characteristics that binding sites might possess compared with the protein surface as a whole. It is important to determine whether the binding sites display an atom compactness and configuration that is distinct from the entire surface distribution. The distribution of inter-triangular distances in binding sites was studied (Figure 4-2). The distances between every pair of triplets (all triangles in the binding site, not only the adjacent ones) in the binding site were calculated. The result was a normal distribution with an average of 10.6Å and a standard deviation of 5.36 Å. The positive skewness (0.96) is graphically demonstrated by the bell curve's long right tail. Exceptionally large binding sites contribute to this skewness by introducing very large inter-triangular distances to the distribution and consequently raising the average distance. This is clear in the right hand side slowly decaying tail of the bell shaped curve. It is impossible to find small counterparts for these high numbers (to restore the symmetry of the distribution) since that would mean having negative distances which is invalid. Nevertheless, the majority of inter-triangular distances in binding sites lie in the region of 2Å to 18Å.

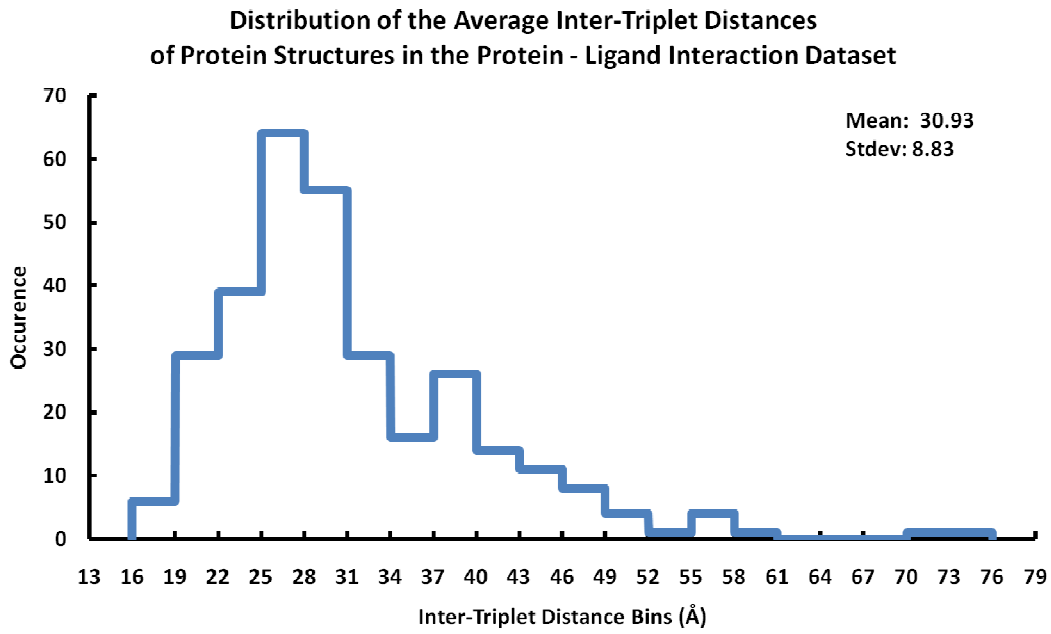

Figure 4-1: Distribution of the average inter-triplet distances on protein surfaces in the protein-ligand interaction dataset.

Comparing the inter-triangular distances between binding sites and the entire protein surfaces faces the problem that protein surfaces are much larger than binding sites surfaces. This skews the protein surface distances towards a larger value (a histogram of average inter-triplet distances on whole protein surfaces is shown in Figure 4-1). Therefore, while studying inter-triangular distances on the protein surfaces, a maximum threshold (5Å) is used to remove this bias. Studying the distances between adjacent triangles was considered (instead of using the 5Å distance threshold). However, the distances between adjacent triangles will not provide a clear measure of the compactness of a binding site as the variations will be minimal. Using a maximum length of 5Å allows for greater variations in the distances to be recorded. Distances between the centroids of any 2 surface triplets (under 5Å) were used to contribute towards the entire protein surface inter-triangular distance average.

Similarly, distances between the centroids of any 2 binding site triplets (under 5Å) were used to contribute towards the entire binding site inter-triangular distance average. The distributions of these 2 values were then compared.

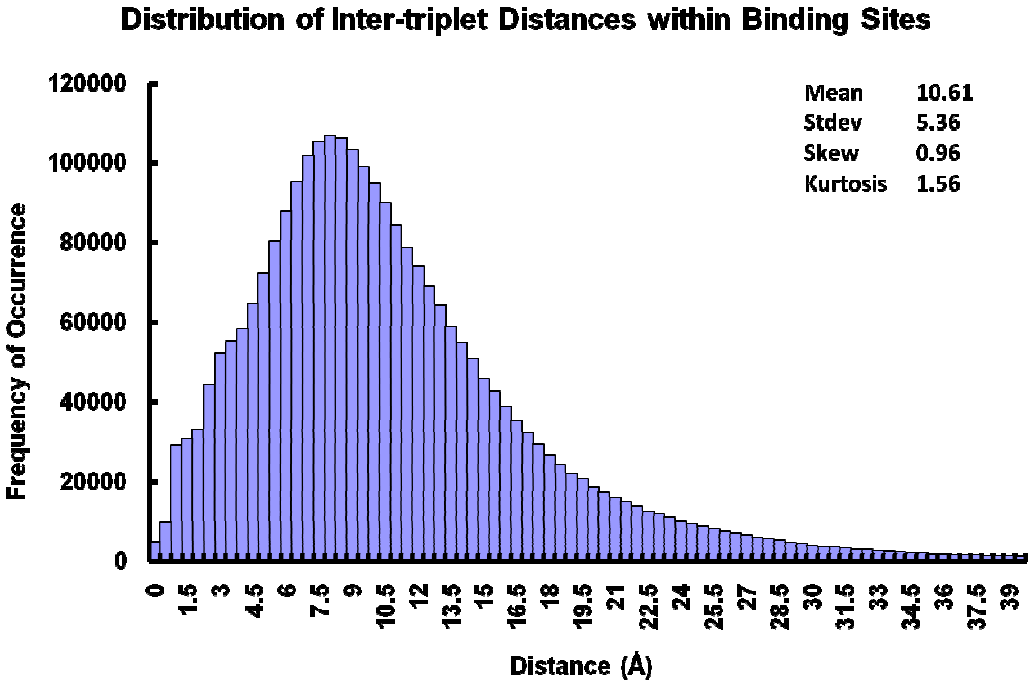

Figure 4-2: The Distribution of Inter-triangular distances between atoms in Binding Sites shows a normal distribution with mean 10.61 Å and standard deviation 5.36 Å. A positive skewness is indicated by the slowly decaying right tail.

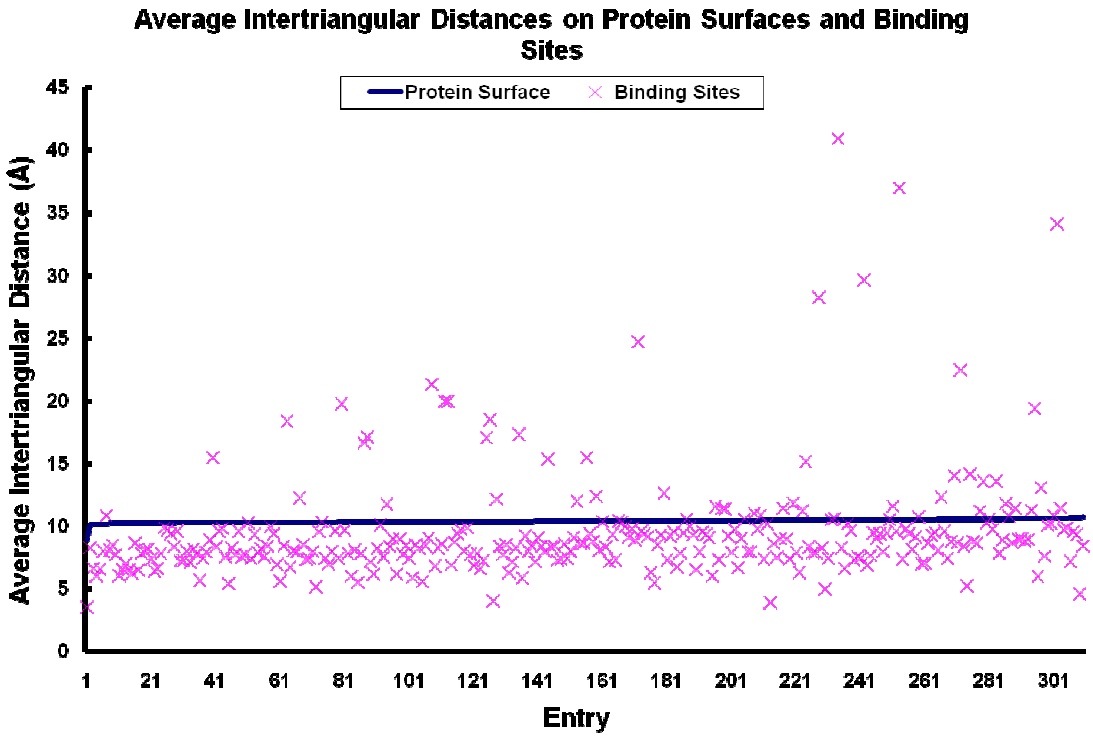

Figure 4-3: The average inter-triangular distances (distance calculated between triangle centroids) measured for all surface triplets (in blue) and all binding site triplets (pink) in the Protein-Ligand Interaction Dataset. Only distances less or equal than 5Å were used for the calculation of the averages. This graph shows that binding sites exhibit a smaller average distance in most cases, indicating that binding sites are more compact. This could be a result of the concavity of the surface in ligand binding sites.

The average inter-triangular distance in binding sites is lower than that of the entire surface in the majority of the cases (Figure 4-3). This property is valid through 90.3% of the protein structures in the dataset, showing that binding sites are generally more compact (by 0.9 Å on average) than the rest of the protein surface. This observation is justified by the topology of binding sites since they comprise a larger number of clefts than the rest of the surface, and this concavity decreases the inter-triangular distances by increasing the number of atoms and triplets within a certain distance around a center atom.

4.1.2 Protein - Protein Interaction Dataset

The distribution of surface and binding site inter-triangular distances in the protein-protein dataset is studied. Inter-triangular distances under 5Å are used to contribute to the average inter-triangular distances in binding site and entire surface distributions (similarly to Section 4.1.1). These distributions are then compared (Figure 4-4), showing that binding sites generally have a lower inter-triangular distance than entire surfaces (in 89% of the cases). However, the difference between these 2 attributes is small (less than 0.1Å) and this is expected since protein-protein interfaces include large patches of the protein surface (unlike protein-ligand and protein-peptide interaction sites which are generally small) and thus are expected to be very similar to the entire protein surface.

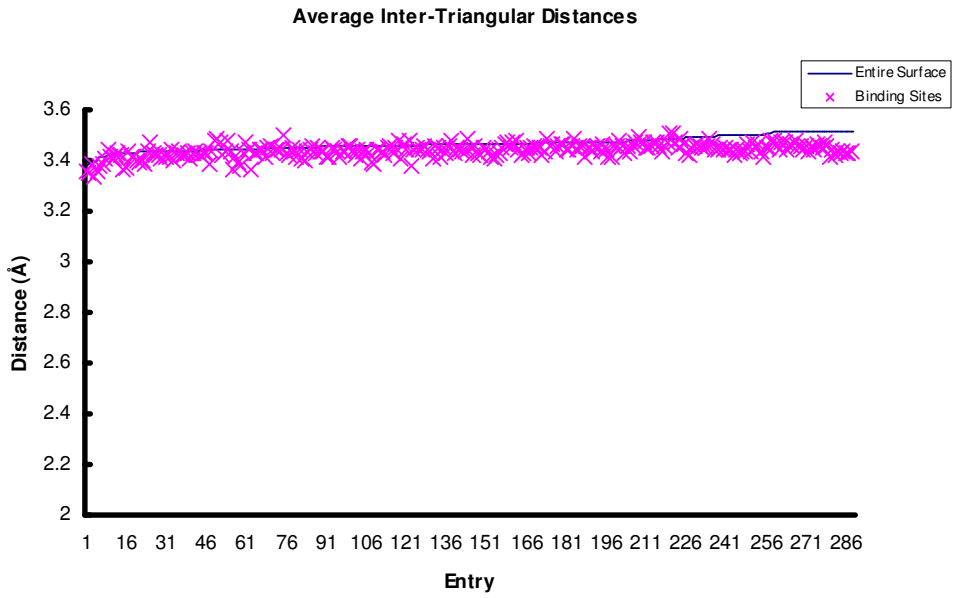

Figure 4-4: The average inter-triangular distances (distance calculated between triangle centroids) measured for all surface triangles (in blue) and all binding site triangles (pink) in the protein-protein Interaction Dataset. Only distances less or equal than 5Å were pooled for the calculation of the averages. Since protein-protein interaction sites are large surfaces and make up a large portion of the entire protein surfaces the 2 distrbutions show similar attributes.

4.1.3 Protein-Peptide Interaction Dataset

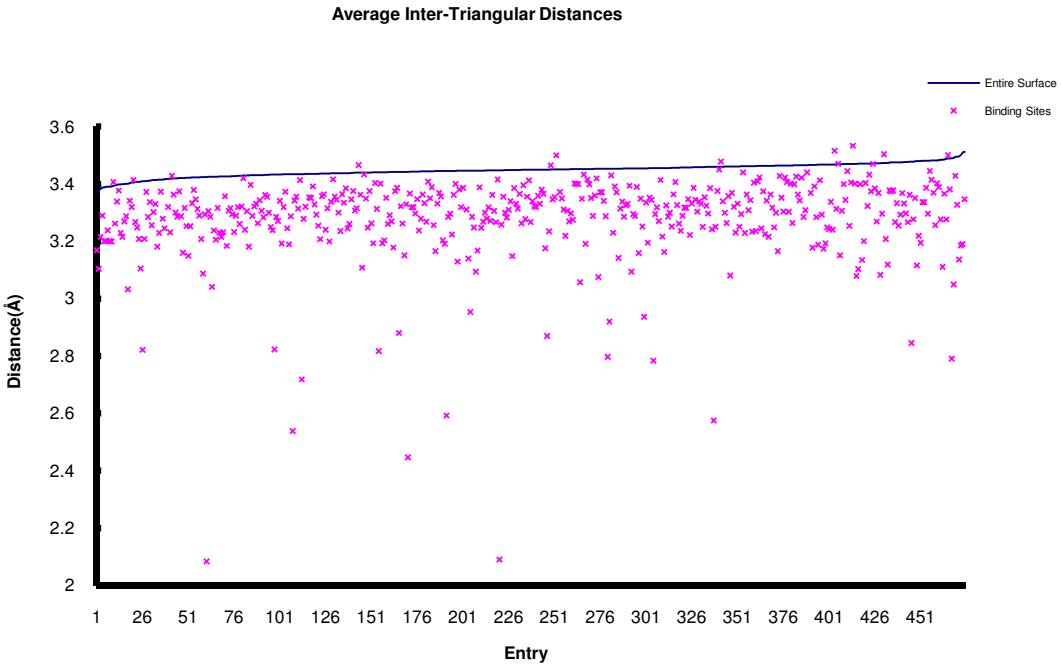

Figure 4-5: The average inter-triangular distances (distance calculated between triangle centroids) measured for all surface triangles (in blue) and all binding site triangles (pink) in the Protein-Peptide Interaction Dataset. Only distances less or equal than 5Å were pooled for the calculation of the averages.

The average inter-triangular distance in binding sites is lower than that of the entire surface in the majority of the cases (Figure 4-5). This property is valid through 97.5% of the protein structures in the dataset, showing that binding sites are generally more compact than the rest of the protein surface. This observation is justified by the nature of protein-peptide interaction. In many cases, this interaction is a result of a peptide sitting in a small pocket or curling around the surface of a certain protein. This ends up in long but thin "rectangular" binding sites which decreases the number of large inter-triangular distances, resulting in a lower binding site average.

4.2 Triangle Areas and Edge Lengths

4.2.1 Protein-Ligand Interaction Dataset

The changes in triplet (triangle) dimensions between binding sites and the entire surface are studied. Figure 4-6 and Figure 4-7 show the distributions of triangle areas and triangle edge lengths in both binding sites and entire surface distributions. Both features are very similar (in terms of average and standard deviation), and that is due to the nature of the triangles being triplets of atoms that come in contact with a water molecule probe simultaneously. This leaves little margin for variability when it comes to the triangle dimensions. In fact, 2 carbon atoms in the same triplet will be furthest apart if the probe falls in between them (having 3 collinear atoms: carbon, probe, carbon; in that order). With a maximum radius of 1.88 Å for a carbon atom (atomic groups C4H1, C4H2, and C4H3), the maximum edge length of a triangle will be 6.56 Å. The minimum edge length will be that of a covalently bonded atoms, and the minimum distance observed in our datasets is 1.14 Å and occurs for atoms Cβ and Oγ of Ser 60 in structure 1BXO.

Although the distributions seem to be similar in general, a few differences exist. The triangle area distributions (Figure 4-6) display a difference in kurtosis. Both curves show negative kurtosis, suggesting that the bell shaped curve is flatter than a standard normal distribution. However, the distribution of binding site triangle areas exhibits a more negative kurtosis, meaning a slightly wider bell-shaped curve.

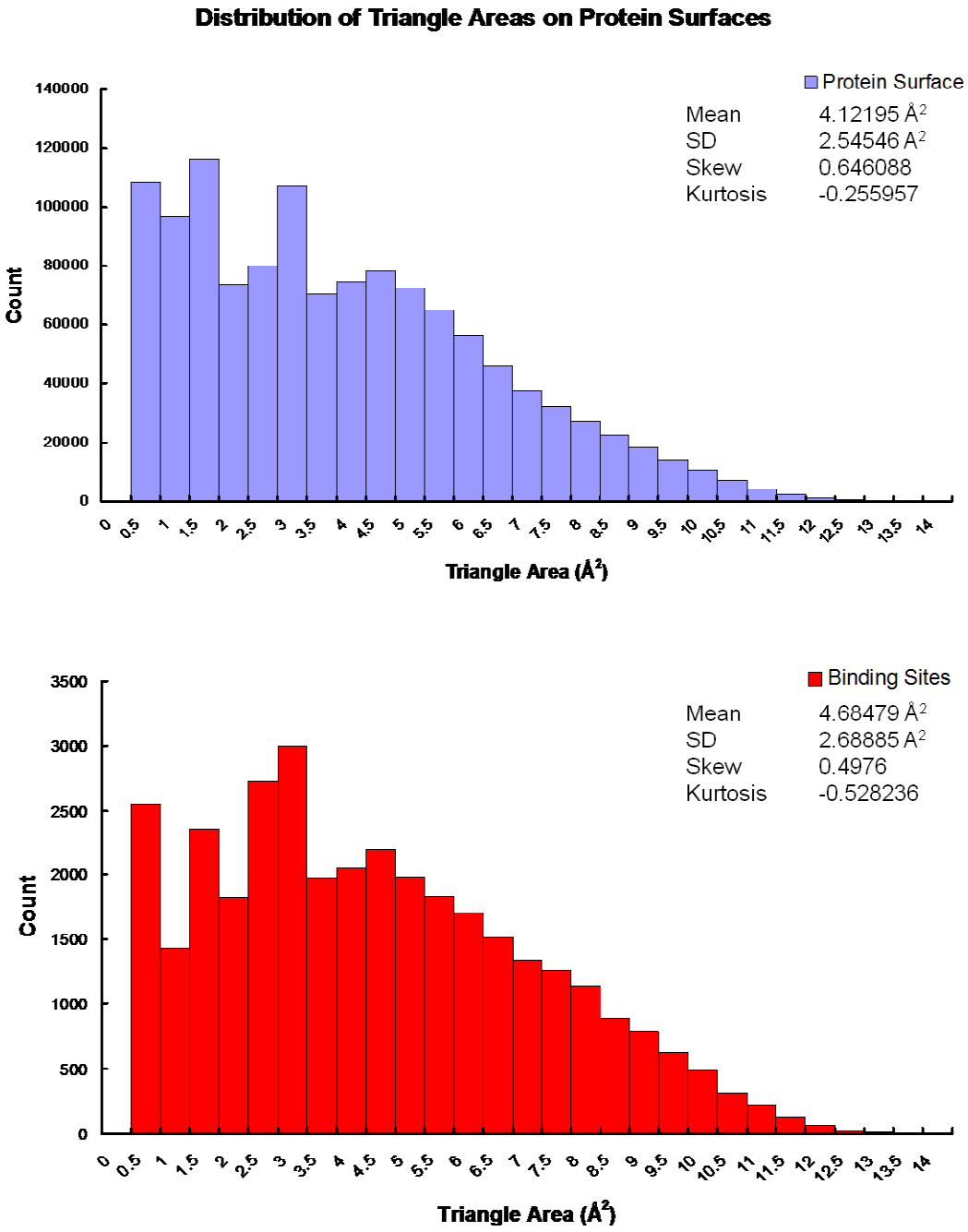

Figure 4-6: Comparison of the distributions of areas of surface and binding sites triangles in the Protein-Ligand interaction dataset shows no distinct differences. The slight difference of the mean of triangle areas is due to the curvature of the binding site surface giving rise to triplets with excentric shapes.

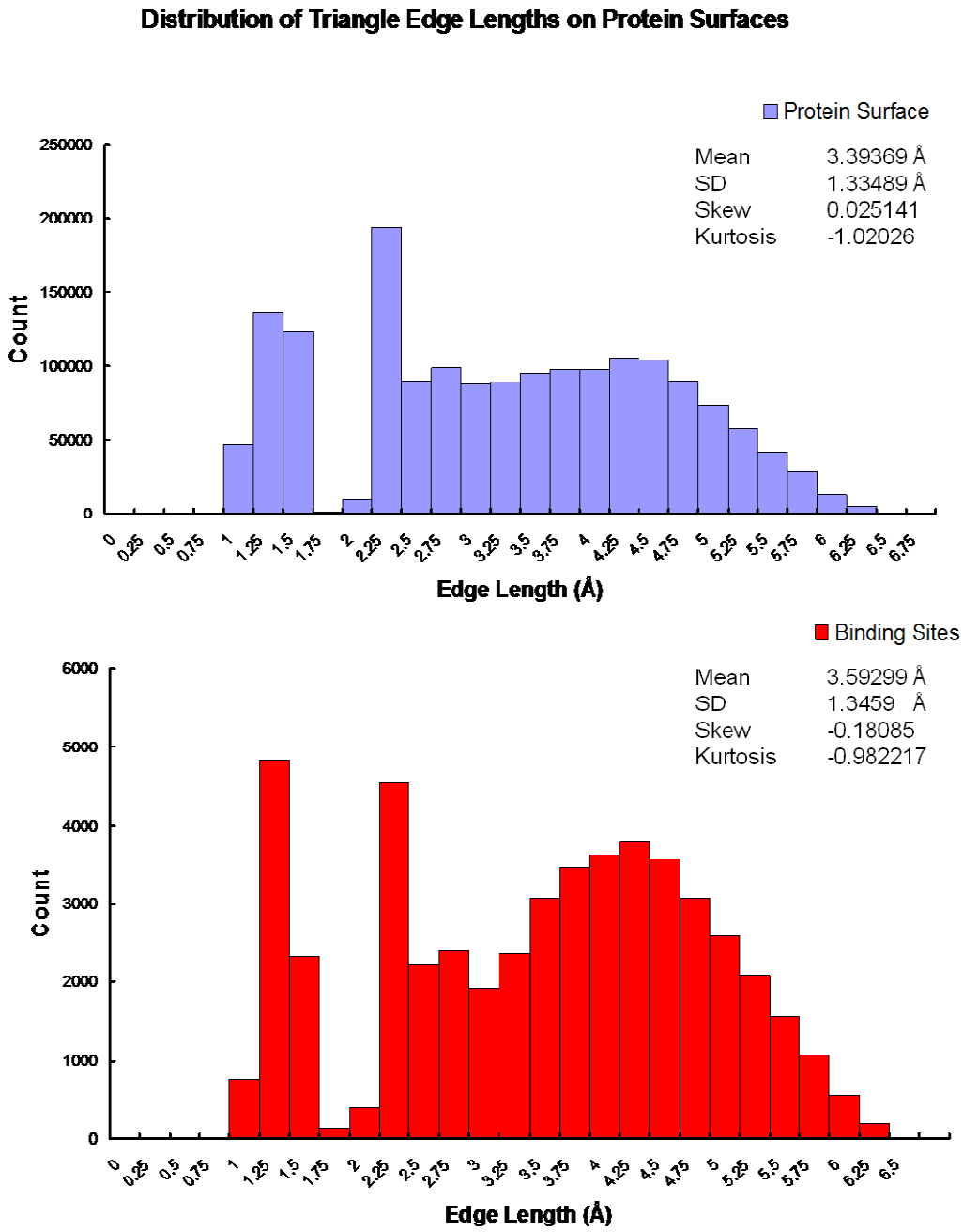

Figure 4-7: Comparison of the distributions of edge lengths of surface and binding sites triangles in the Protein-Protein interaction dataset shows no distinct differences.

The edge length distribution (Figure 4-7) shows more interesting differences. The Entire surface distribution has virtually zero skewness, but the binding sites distribution exhibits a slightly negative skewness. This is due to the distances between 1Å and 1.75Å contributing more to the binding sites distribution than to the entire surface. This atom compactness in binding sites has already been shown in Section 4.1, and can also be justified by the amino acid side chains pointing outwards in binding sites since they generally play an important role in binding. The exposure of these side chains to the water surface leads to an increased number of these small distances. The triangle edge lengths in binding sites show a significant increase in the regions of 1Å to 1.75Å and 3.75Å to 5Å when compared to the entire surface distribution. This observation, put together with the fact that the average triangle areas does not change brings forward a conclusion that in binding sites, triangles with shorter and longer sides occur more frequently. This could be a sign of increased concavity, giving rise to triangles with acute shapes. A second high peak appears at [2.25Å -2.5Å]. Such distances are recorded for acidic or aromatic systems like OE1 and NE2 of Gln, NH1 and NH2 of Arg, and CE2 and CE3 of Trp.

4.2.2 Protein-Protein Interaction Dataset

The areas and edge lengths of surface and binding site triangles in the protein-protein dataset is studied (similar to Section 4.2.1). Although the distributions of triangle areas in binding sites and protein surfaces are similar (Figure 4-8), binding site triangles exhibit a smaller area (depicted by the mean area for binding site triangles being smaller than the mean area of surface triangles and also by the shifting of the distribution histogram to the left). This is also reflected in the distributions of triangle edge lengths in binding sites and protein surfaces (Figure 4-9). A slight increase in

triangle edges between 1.25Å and 1.5Å is noted. This could be a result of the induced fit both proteins undergo when they bind to each other at large patches of the surface, increasing the compactness of some parts of the binding surface. This does not result in a major change in the inter-triangular distances (Figure 4-4) as a result of the large number of triangles in protein-protein interfaces, which renders the impact of a change in the geometry of some interface triangle less influential on the entire interface distribution.

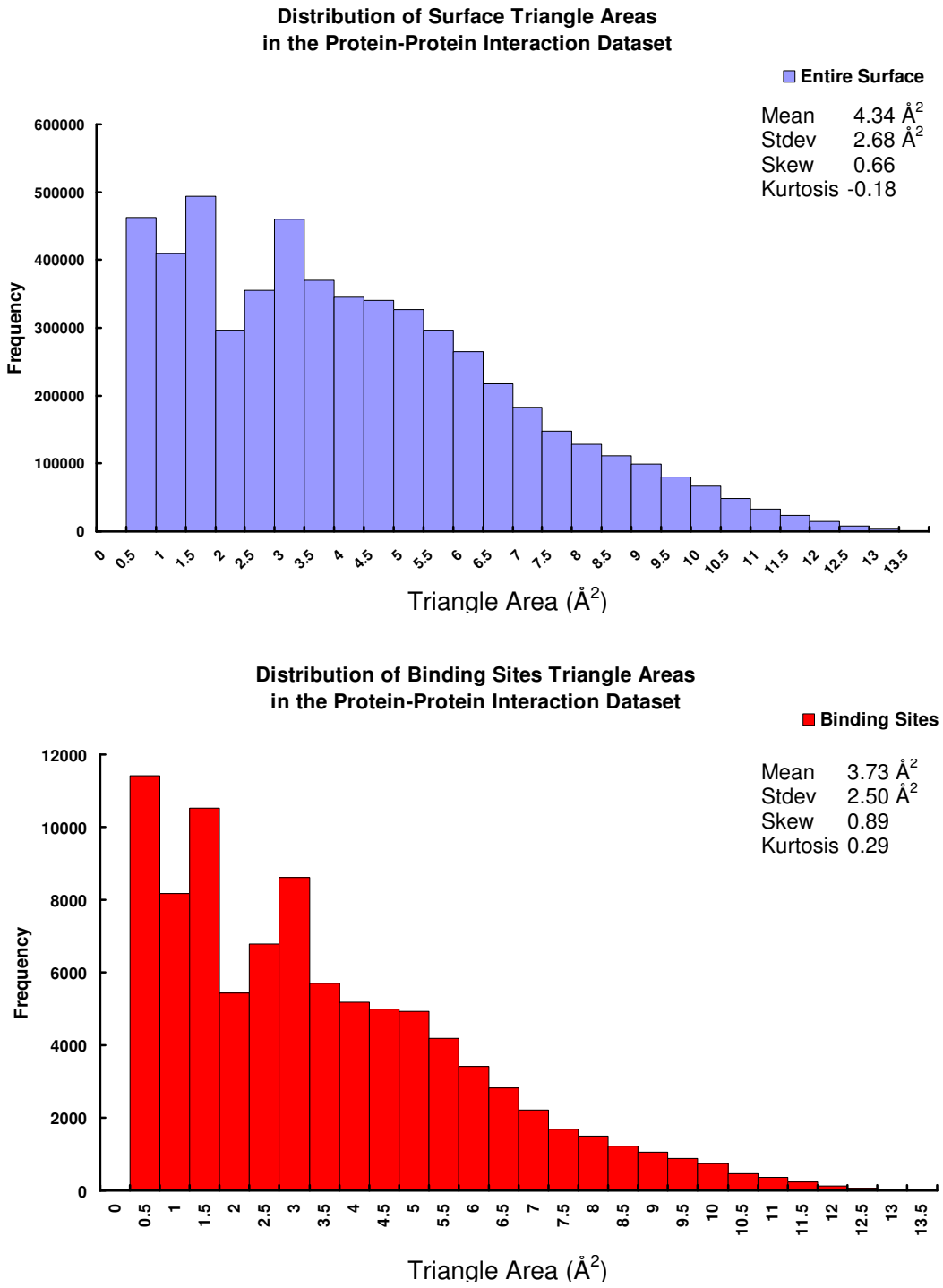

Figure 4-8: Comparison of the distributions of areas of surface and binding sites triangles in the Protein-Protein interaction dataset shows no distinct differences.

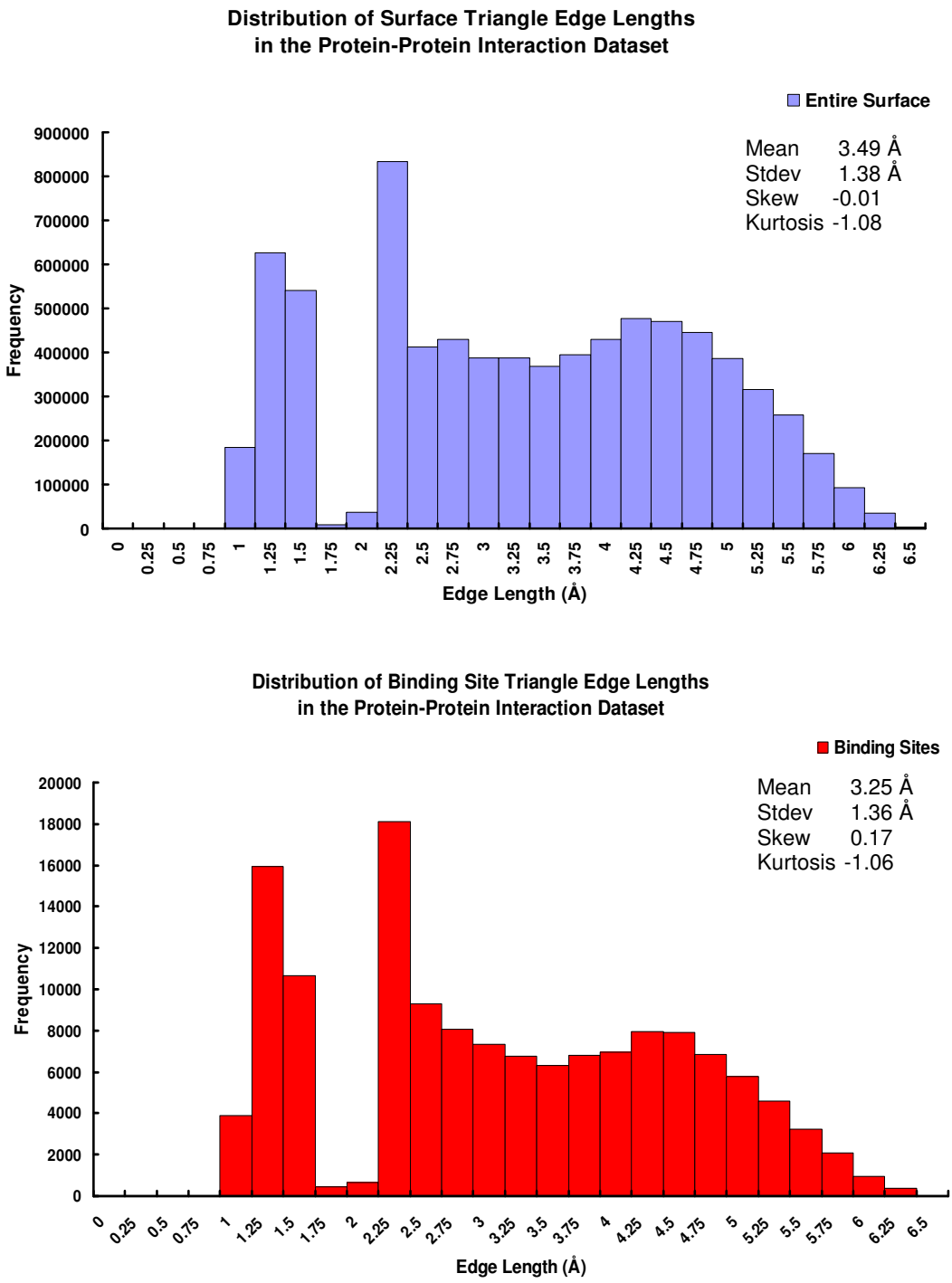

Figure 4-9: Comparison of the distributions of edge lengths of surface and binding sites triangles in the Protein-Protein interaction dataset shows no distinct differences.

4.2.3 Protein-Peptide Interaction Dataset

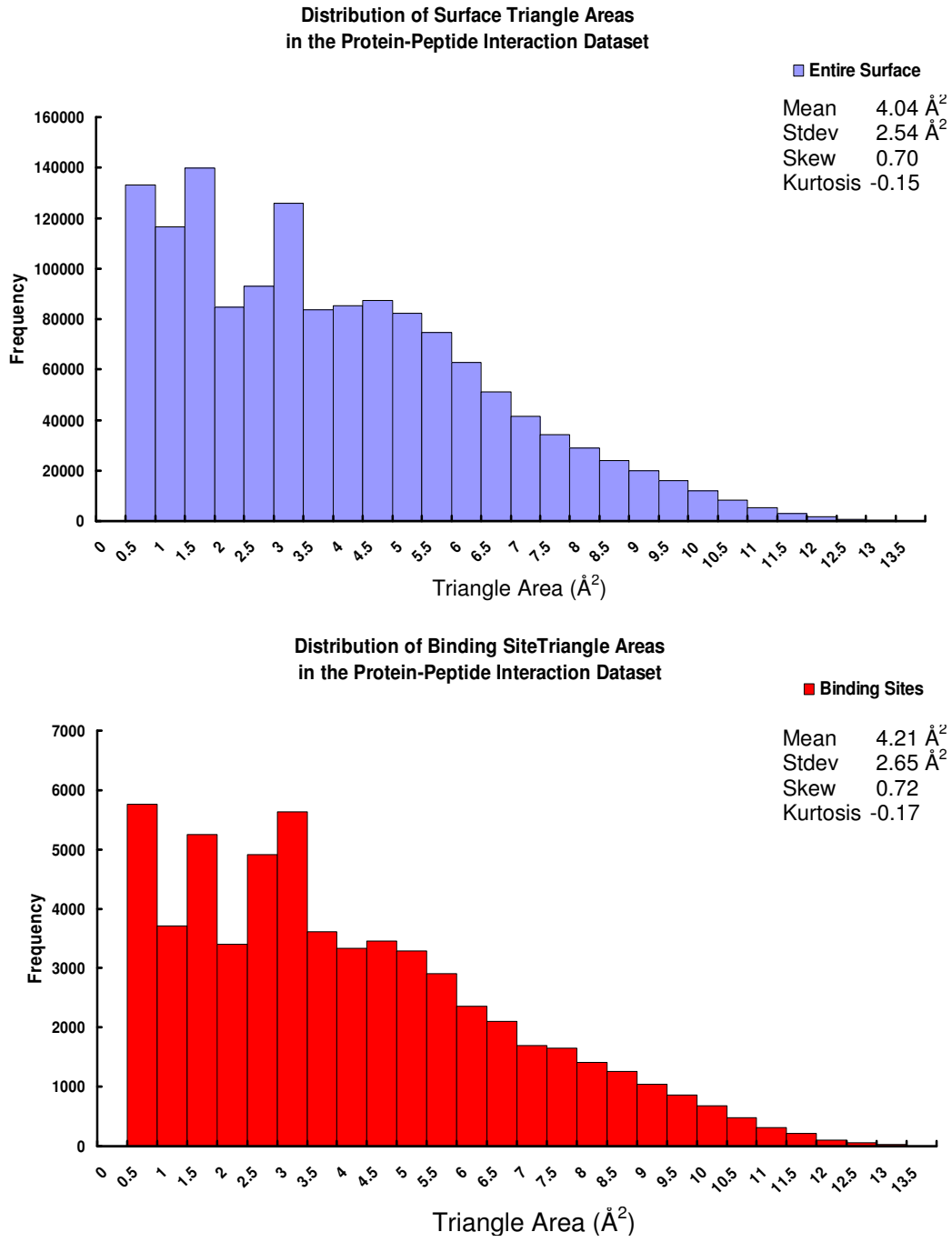

Figure 4-10: Comparison of the distributions of areas of surface and binding sites triangles in the Protein-Peptide interaction dataset shows no distinct differences.

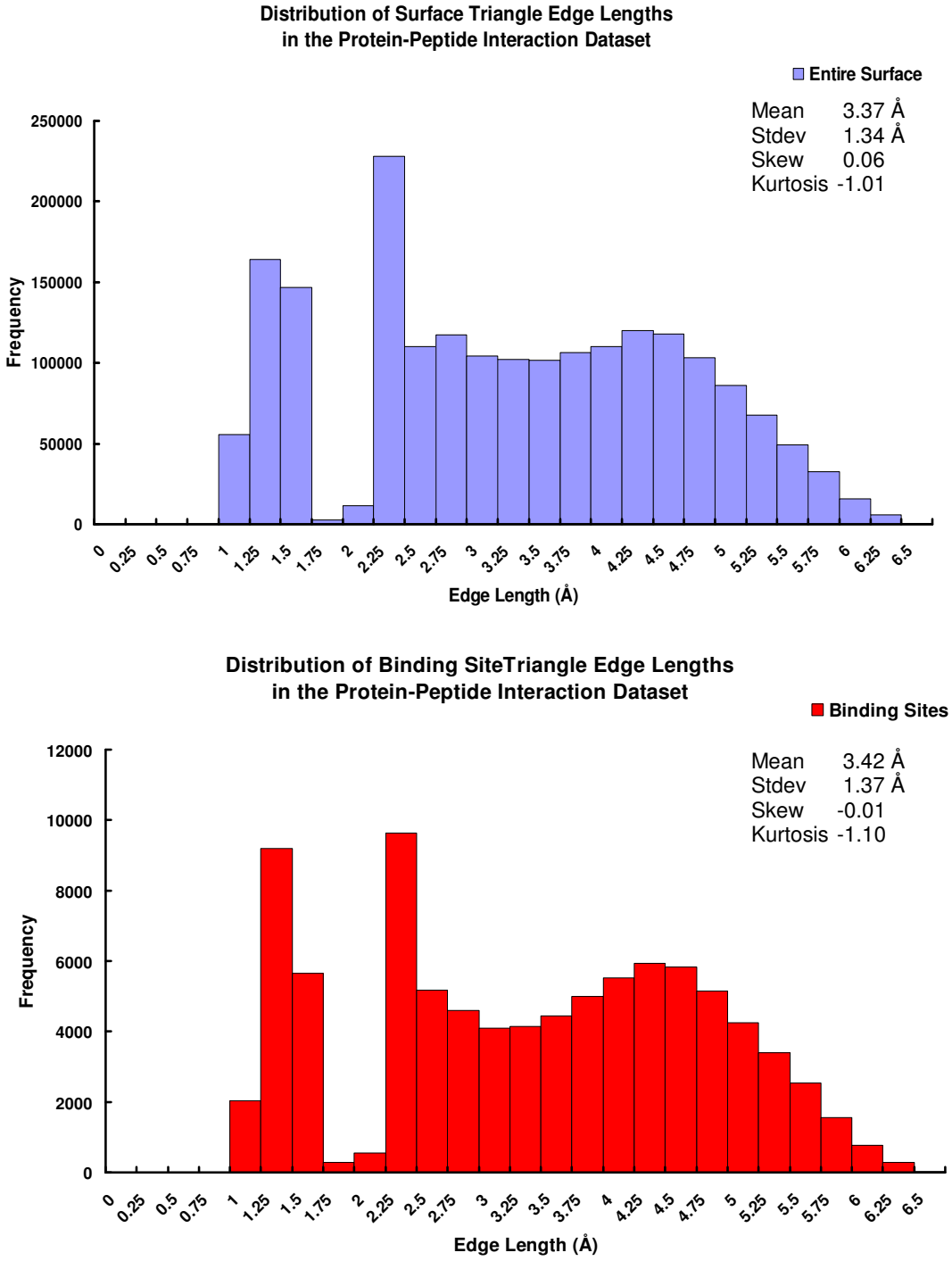

Figure 4-11: Comparison of the distributions of edge lengths of surface and binding sites triangles in the Protein-Peptide interaction dataset shows no distinct differences.

The areas and edge lengths of surface and binding site triangles in the protein-peptide dataset is studied (similar to Section 4.2.1). The distributions of triangle areas in binding sites and protein surfaces (Figure 4-10) are similar in all aspects (average, standard deviation, skewness, and kurtosis). This is also reflected in the distributions of triangle edge lengths in binding sites and protein surfaces (Figure 4-11). A slight increase in triangle edges between 1.25Å and 1.5Å and also between 4Å and 4.75Å is noted (contributed by acute triangle shapes that exist in cavities). These differences are not large enough to induce a large change in the distribution of triangle areas and edge lengths distributions.

4.3 Spatial Differences between Binding Sites of the Ligand, Peptide, and the Protein Interaction Datasets

Comparing the binding sites of the three interaction types may provide insight into the general topology of each type of interaction. Ligands generally bind in deep semispherical pockets, peptides bind on the surface in long thin strips, and proteins bind each other at large surface patches. This may have implications on the inter-triangular distances, triangle areas, and edge lengths. However, differences in the inter-triangular distances are more pronounced than triangle areas and edge lengths for the different binding site topologies. Due to the nature of the triangles being triplets of surface atoms that are touched by a probe sphere at the same time, the variation in the shape of the triangles is very limited; the space in which those three atoms can exist whilst still being in contact with the probe sphere and without any steric clash with each other and with other protein atoms is small.

Inter-triangular distances show distinct properties across the binding sites of the three interaction classes. Protein-Protein binding sites have an average inter-triangular distance that is very close (but still less) to that of the entire surface. This similarity is due to the large binding sites in which protein-protein interaction takes place. These binding sites include cavities of different sizes in addition to some flat surfaces as well (just like a normal protein surface). Protein-Ligand and Protein-Peptide binding sites have smaller inter-triangular distances (3.25Å and 3.33Å respectively compared to 3.43Å for protein-protein interactions). This difference is significant as this value is an average of all distances under 5Å. This is a direct result of the binding taking place in deep pockets where the surface is concave, including more triangles within the same distance of a certain central triangle. The deep pockets are sometimes very deep and cross from one side of the protein to the other, resulting in cases with average inter-triangular distances higher than that of the entire surface. In contrast, Protein-Peptide binding sites have the least percentage of cases with inter-triangular distances higher than that of the entire surface. This is due to the rectangular shape of peptide binding sites as this decreases the number of inter-triangular distances above a certain limit, leading to the decrease of the average inter-triangular distance.

4.4 Chemical Composition of Triplets

The Protein-Ligand dataset (made up of 309 structures) generated 1.223 million surface triplets out of which 34,228 were binding site triplets. The Protein-Peptide dataset (made up of 475 structures) generated 1.416 million surface triplets out of which 59,145 were binding site triplets. The Protein-Protein dataset (made up of 210 structures) generated 1.708 million surface triplets out of which 93,054 were binding site triplets. The large increase in the number of triplets (relative to the number of

structures) for the Protein-Protein dataset is due to the interaction occurring between two large globular domain which increases the number of both surface and interface triplets. This section studies the chemical composition of the binding sites and protein surfaces in each of these datasets.

The triplets are sub classified into four categories (Table 2): There are a total of 35 types of 'hydrophobic triplets' containing three hydrophobic (carbon) atoms. The 120 types of 'polar triplets' consist of permutations of three polar (N,O,S) atoms. The 120 types of 'mostly hydrophobic triplets' contain two hydrophobic atoms and the 180 types of 'mostly polar triplets' contain two polar atoms (Table 4-1).

Table 4-1: The classification of the STP surface triplets by chemical composition

Triplet Type	Chemical Composition	Number of Triplets	Ratio of all Triplets
Hydrophobic	3 Carbon Atoms	35	0.08
Mostly Hydrophobic	2 Carbon Atoms	120	0.26
Mostly Polar	1 Carbon Atom	180	0.40
Polar	Zero Carbon Atoms	120	0.26

STP is created with three versions that study Protein-Ligand (Section 2.3.1), Protein-Peptide (Section 2.3.2), and Protein-Protein (Section 2.3.3) interactions. The triplets in each of these datasets were studied based on the chemical classification described above (Table 4-1). Several attributes were analyzed including the range, average, and standard deviation of the propensity scores of the triplets that make up these sub-categories (Table 4-2). The Protein-Peptide and Protein-Protein datasets show similar results. The interaction between protein domains is in fact between the peptide stretches that form their binding interfaces, and the similarity of propensity scores in

both profiles concurs with this fact. 'Polar triplets' have the highest average propensity score in the Protein-Ligand dataset while 'hydrophobic triplets' have the highest average propensity score in the Protein-Peptide and Protein-Protein datasets (Table 4-2). This is due to the ligands in the Protein-Ligand dataset being much smaller than the peptides or proteins in the other two datasets, suggesting a more important role for polar interactions in contrast to the dominance of hydrophobic interactions when it comes to the interaction between large domains. This is also evident in 'mostly polar triplets' having a higher average propensity than 'mostly hydrophobic triplets' in the Protein-Ligand dataset while the latter have higher average propensities in the Protein-Peptide and Protein-Protein Datasets. All triplet sub-categories in all datasets exhibit similar standard deviation of the propensity scores.

Table 4-2: The distribution of propensities for different triplet types (Table 4-1) in the protein-ligand, protein-peptide, and protein-protein dataset.

Interaction Dataset	Triplet Type	Propensity Range	Propensity Average, standard deviation
Protein - Ligand	Hydrophobic	-1.3, 2.7	0.7, 1.1
	Mostly Hydrophobic	-2.2, 3.5	0.4, 1.2
	Mostly Polar	-3.5, 5.2	0.7, 1.3
	Polar	-0.4, 4.2	1.4, 1.2
Protein - Peptide	Hydrophobic	-1.1, 2.2	0.8, 0.9
	Mostly Hydrophobic	-1.8, 2.9	0.4, 1.0
	Mostly Polar	-2.7, 3.0	0.2, 1.0
	Polar	-1.7, 4.6	0.4, 1.3
Protein - Protein	Hydrophobic	-0.8, 2.1	0.7, 0.7
	Mostly Hydrophobic	-2.7, 4.1	0.2, 1.1
	Mostly Polar	-4.4, 4.1	-0.1, 1.4
	Polar	-2.6, 3.1	-0.4, 1.2

We next study the occurrence of these triplet sub-categories. Hydrophobic atoms constitute 61% of the surface in each of the 3 interaction datasets, and Polar atoms constitute 39%. Therefore, a 'Hydrophobic triplets' has an *expected occurrence rate* of 0.61^3, and a 'polar triplet' has an expected occurrence rate of 0.39^3. 'Mostly hydrophobic and mostly polar triplets' have *expected occurrence rates* of ($0.61^2 \times 0.39 \times 3$) and ($0.39^2 \times 0.61 \times 3$) respectively. This ratio is calculated for all triplet subcategories (Table 4-1) and an expected occurrence per dataset is calculated as the *expected occurrence rate* of a triplet category multiplied by the total number of triplets in a dataset. This expected occurrence is compared with the actual occurrence in binding sites and entire surfaces for the three interaction datasets (Table 4-3 and Table 4-4).

Table 4-3: The occurrence of different triplet subcategories on the protein surface of the structures in the protein-ligand, protein-peptide, and protein-protein dataset. The expected occurrence is calculated as the fraction of a specific type of triplet to all triplet types multiplied by the total number of triplets in the dataset. The occurrence ratio is calculated by dividing the actual occurrence by the expected occurrence.

Interaction Dataset	Triplet Type	Occurrence in dataset (A)	Expected Occurrence (B)	Occurrence Factor (A ÷ B)
Protein - Ligand	Hydrophobic	204,987	277,600	0.74
	Mostly Hydrophobic	608,379	532,445	1.14
	Mostly Polar	360,230	340,416	1.06
	Polar	49,412	72,548	0.68
Protein-Peptide	Hydrophobic	250,873	321,510	0.78
	Mostly Hydrophobic	707,095	616,666	1.15
	Mostly Polar	406,941	394,262	1.03
	Polar	51,551	84,023	0.61
Protein - Protein	Hydrophobic	309,548	387,870	0.8
	Mostly Hydrophobic	846,557	743,948	1.14
	Mostly Polar	486,885	475,639	1.02
	Polar	65,812	101,366	0.65

Table 4-4: The occurrence of different triplet subcategories in the binding sites of the structures in the protein-ligand, protein-peptide, and protein-protein dataset. The expected occurrence is calculated as the fraction of a specific type of triplet to all triplet types multiplied by the total number of triplets in the dataset. The occurrence ratio is calculated by dividing the actual occurrence by the expected occurrence.

Interaction Dataset	Triplet Type	Occurrence in dataset (A)	Expected Occurrence (B)	Occurrence Factor (A ÷ B)
Protein - Ligand	Hydrophobic	8,883	7,783	1.14
	Mostly Hydrophobic	14,218	14,928	0.95
	Mostly Polar	9,306	9,544	0.98
	Polar	1,881	2,034	0.92
Protein-Peptide	Hydrophobic	17,486	13,425	1.3
	Mostly Hydrophobic	26,918	25,749	1.05
	Mostly Polar	13,082	16,463	0.79
	Polar	1,659	3,508	0.47
Protein - Protein	Hydrophobic	25,507	21,121	1.21
	Mostly Hydrophobic	45,182	40,512	1.12
	Mostly Polar	20,362	25,902	0.79
	Polar	2,003	5,520	0.36

The expected occurrence and the actual occurrence of the triplet sub-categories were compared by dividing the actual occurrence by the expected occurrence. We call this attribute the “Occurrence Factor”. An Occurrence Factor greater than 1 would indicate and over expressed sub-category while an Occurrence Factor less then 1 would indicate and under expressed sub-category. The three interaction datasets exhibit similar occurrence factors for the triplet sub-categories all over the surface (Table 4-3). ‘Mostly hydrophobic and mostly polar triplets’ are overexpressed while ‘hydrophobic and polar’ triplets are under expressed. However, a different result is observed when the occurrence factors are calculated for binding site triplets (Table 4-4). Comparing the results in Table 4-3 and Table 4-4 shows that the Occurrence Factors for ‘hydrophobic triplets’ is higher in binding sites than it is on the entire

protein surface. This is expected since hydrophobic interactions play important roles in protein function. Interestingly, the occurrence factor of 'polar triplets' in binding sites is larger than all over the surface in the sole case of Protein-Ligand interaction, while the opposite is observed in the case of Protein-Peptide and Protein-Protein interaction.

$$\text{Triplet Class Propensity}(\alpha) = \frac{\text{InterRatio}(\alpha)}{\text{SurfaceRatio}(\alpha)}$$

Equation 4-1: Calculation of the Triplet Class Propensity for the different triplet subcategories. InterRatio(α) corresponds to the occurrence factor of triplet type α in the interface (Table 4-4). SurfaceRatio(α) corresponds to the occurrence factor of triplet type α on protein surfaces (Table 4-3).

Table 4-5: The actual occurrence ratios and triplet class propensity for the triplet sub-categories. The actual occurrence ratios are the result of a division of the number of occurrences of a certain triplet by the total number of triplets. The Triplet Class Propensity is a result of dividing the Binding Site Actual Occurrence Ratio by the Entire Surface Actual Occurrence Ratio (Equation 4-1).

Interaction Dataset	Triplet Type	Binding Site Occurrence Factor (A)	Entire Surface Occurrence Factor (B)	Triplet Class Propensity (A ÷ B)
Protein - Ligand	Hydrophobic	1.14	0.74	1.54
	Mostly Hydrophobic	0.95	1.14	0.83
	Mostly Polar	0.98	1.06	0.92
	Polar	0.92	0.68	1.35
Protein-Peptide	Hydrophobic	1.3	0.78	1.67
	Mostly Hydrophobic	1.05	1.15	0.91
	Mostly Polar	0.79	1.03	0.77
	Polar	0.47	0.61	0.77
Protein - Protein	Hydrophobic	1.21	0.8	1.51
	Mostly Hydrophobic	1.12	1.14	0.98
	Mostly Polar	0.79	1.02	0.77
	Polar	0.36	0.65	0.55

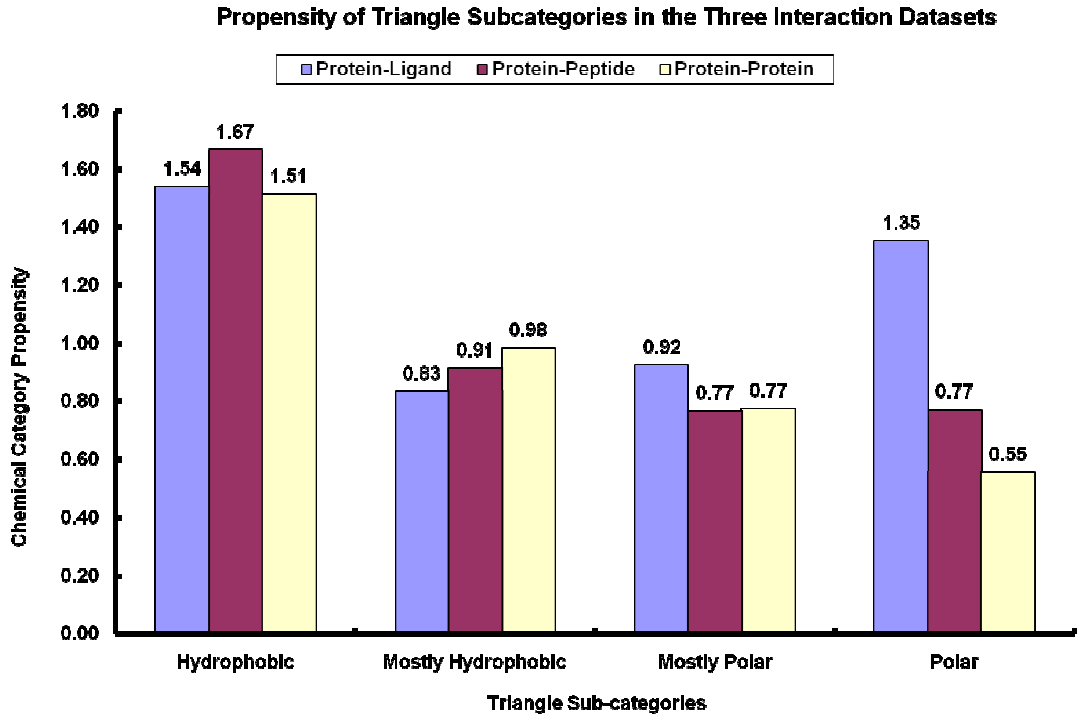

Figure 4-12: The Triplet Class Propensity for the triplet subcategories in the three interaction datasets. This propensity is calculated as ratio binding site triplets that are of a certain category divided by the ratio of all surface triplets that are of that same category (Equation 4-1).

Comparing the occurrence of the subcategories between binding sites and entire protein surfaces can be better conveyed using the Triplet Class Propensity (TCP) scores (Equation 4-1). The Triplet Class Propensity compares the occurrence factor of a certain sub-category of triplets in the binding site (Table 4-4) with the occurrence ratio of that category all over the surface (Table 4-3). This Triplet Class Propensity is reported in Table 4-5 and Figure 4-12. 'Hydrophobic triplets' play an important role in the interaction between proteins and their protein, ligand, or peptide binding partners. 'Polar triplets' however play an important role in the interaction of Protein-Ligand complexes only. This study shows that the role of "Polar" atoms increases in Protein-Ligand interactions, signaled by the increase in the Triplet Class

Propensity of ‘polar and mostly polar triplets’ and the decrease of the Triplet Class Propensity of ‘mostly hydrophobic triplets’ (Figure 4-12).

4.5 Recognition of certain atoms by specific Triplets

Table 4-6: The Classification of the Ligand Atom Types [214] in the Protein-Ligand interaction dataset.

Atom	Classification
Br	Halogen
Cl	Halogen
F	Halogen
F.3	Halogen
C.1	Hydrophobic
C.2	Hydrophobic
C.3	Hydrophobic
C.ar	Hydrophobic
N.ar	Hydrophobic
N.1	Polar
N.2	Polar
N.3	Polar
N.am	Polar
N.pl3	Polar
P.3	Polar
S.2	Polar
S.3	Polar
S.o2	Polar
O.2	Polar
O.3	Polar
NON	Empty
HOH	Water

The triangular types proved useful in predicting the location of binding sites, classifying enzyme types, and ranking docking orientations. We now study the interaction between the surface triplets and ligand atoms, and search for preferences that some triplets might have for different ligand atom types. Ligand atoms in the dataset are classified as per the Tripos [214] forcefield definitions. The protein-

ligand dataset contained 21 ligand atom types (Table 4-6). To simplify the problem of studying what triplets interact with what atoms, ligand atoms were classified into the 4 classes: Halogens, Hydrophobic, Polar, and Water.

Table 4-7: The recognition of various ligand atom classes by the triplet classes in the protein – ligand interaction dataset shows a strong affinity between 'hydrophobic triplets' and Hydrophobic atoms, 'polar triplets' and Polar atoms, and 'hydrophobic triplets' and Halogens. OF/EF values can be transformed into statistical free energy by the formula $\Delta G_{stat} = -RT \times \ln(OF/EF) / 1000$ to give a value in kcal/mol.

Frequency	Triplet Class / Atom Class	Hydrophobic	Mostly Hydrophobic	Mostly Polar	Polar	Total
Observed Frequencies (OF)	Empty	58696	148453	75323	8430	290902
	Halogen	93	73	32	8	206
	Hydrophobic	5545	6023	2577	362	14507
	Polar	1810	4101	3406	734	10051
	Water	138843	449729	278892	39878	907342
	Total	240987	608379	360230	49412	1223008
Expected Frequencies (EF)	Empty	48758	144708	85684	11753	290902
	Halogen	35	102	6	8	206
	Hydrophobic	2432	7216	4273	586	14507
	Polar	1685	5000	2960	406	10051
	Water	152079	451353	267252	36658	907342
Interaction Preference OF/EF	Empty	1.20	1.03	0.88	0.72	N/A
	Halogen	2.69	0.71	0.53	0.96	N/A
	Hydrophobic	2.28	0.83	0.60	0.62	N/A
	Polar	1.07	0.82	1.15	1.81	N/A
	Water	0.91	1.00	1.04	1.09	N/A
$\Delta G_{stat} = -RT \times \ln(OF/EF) / 1000$ (kcal/mol)	Empty	-0.108	-0.018	0.076	0.195	N/A
	Halogen	-0.586	0.203	0.376	0.024	N/A
	Hydrophobic	-0.488	0.110	0.303	0.283	N/A
	Polar	-0.040	0.118	-0.083	-0.352	N/A
	Water	0.056	0.000	-0.023	-0.051	N/A

The interaction between triplets and ligand atoms is then quantified as follows: for each surface triplet, the closest (distance between atom center and triplet centroids) ligand atom or water molecule is recorded as an interaction partner. If there are no atoms within a distance of 4Å, the closest atom type is recorded as "Empty". The interaction between each triplet Class (Hydrophobic, Mostly Hydrophobic, Mostly

Polar, and Polar) and each atom class (Halogen, Hydrophobic, Polar, and Water, Empty) is studied. The observed frequencies of each interaction are recorded and compared with the expected frequency. The expected frequency of a certain interaction depends on the availability of a certain triplet class and a certain atom class in the dataset. For example, if we have 206 halogen atoms in the database, and 240,987 hydrophobic triplets, and a total number of interactions of 1,223,008, the expected frequency of Hydrophobic/Halogen interactions is 206 × 240,987 ÷ 1,223,008. After the expected frequencies are calculated, a final attribute is calculated by dividing the Observed Frequency by the Expected Frequency of a certain interaction. A result greater than 1 indicates the favouring of a certain interaction while a result less than 1 indicates a certain interaction being disfavoured (Table 4-7).

The triplet:ligand atom interaction data indicated a higher tendency for 'hydrophobic and mostly hydrophobic triplets' to have no binding partners. These triplets are abundant all over the surface (Table 4-3) and have a lower propensity of interaction to water (as they are hydrophobic) and thus are rendered with no binding partners in the crystal structures. In contrast, 'mostly polar and polar triplets' attract more water molecules than 'hydrophobic and mostly hydrophobic triplets' and have less affinity to be without a binding partner in the crystal structure (as a water molecule is attracted to them). Hydrophobic ligand atoms are attracted to 'hydrophobic triplets', but not to "mostly hydrophobic triplets', as they are possibly repelled by the polar atom in these triplets. However, the tendency of a hydrophobic atom to interact closely with 'mostly hydrophobic triplets' is higher than its tendency to interact with

a ‘mostly polar or polar triplets’ (Table 4-7). Halogens show a distinct affinity towards ‘Hydrophobic triplets’ compared with everything else. However, the Halogen atom : ‘polar triplet’ interaction is preferred over the Halogen atom / Mostly Hydrophobic interaction This is a key indicator in the organohalogens’ dual nature as hydrogen bond acceptors that fit comfortably in hydrophobic environments [215].

According to the Boltzmann and Gibbs classification of free energy [216, 217], a reaction of type A + B → C would have a statistical free energy according to the following equation:

$$free\ energy\ \Delta G_{stat} = -R \times T \times \ln \frac{P(C)}{P(A) \times P(B)}$$

where P(A), P(B), and P(C) are the probabilities of finding these substances in solution, i.e. the concentration of these substances.

Equation 4-2: The free energy of a chemical interaction

If we consider that a Triplet T reacts with a ligand atom A to produce and interacting complex C, the reaction would be of the form T + A → C. The value of the expression P(C) / [P(T) x P(A)] as given by the Gibbs free energy formula has already been computed as the interaction preference in Table 4-7. Hence, the interaction preference of the different triplets with particular atom types can be used to give a measure of statistcal ΔG values (Equation 4-2) according to the formula ΔG_{stat} = -R × T × ln (interaction preference).

For example, a non-bonded interaction between a Halogen-class atom and a 'hydrophobic triplet' has an interaction preference of 2.7, where the interaction preference is calculated as the Observed Frequency (OF) / Expected Frequency (EF) (Table 4-7). This is equivalent to saying that the frequency of interaction of a halogen atom with a 'hydrophobic triplet' is about three times the expected value. The statistical free energy difference that accounts for this distribution can be calculated from: $\Delta G_{stat} = -RT \ln (OF / EF)$, where R is the gas constant = 1.9872 cal deg^{-1} mol^{-1}. This gives a ΔG_{stat} = -0.59 kcal/mol at 298K for the interaction of Halogen atoms with 'hydrophobic triplets'. The other clear preference for atom environment (Supplementary Table 6) is the interaction between Hydrophobic atoms and 'hydrophobic triplets' (interaction preference 2.28) which gives a ΔG_{stat} of -0.49 kcal/mol. The preference of Polar atoms interacting with 'polar triplets' is less marked with a ΔG_{stat} of -0.35 kcal/mol.

4.6 STP Propensities and Statistical Free Energy Values

The propensities of the 455 triplet types (defined in Chatper 1, Equation 2-1) range between -3.54 and 5.16. Similarly to section 4.5, these propensity values can also be related to statistical free energy values. Since the propensity is the $\log_2$(Interface Probability / Surface Probability), getting the probabilities back is generated be retrieved as (propensity)×ln(2). Thus, the statistical free energy calculations for the occurrence of a triplet type in binding sites can be calculated as:

$$\Delta G_{stat} = - R \times T \times \ln(2) \times \text{propensity}$$

Equation 4-3: Transformation of STP propensities into statistical free energy values

For the protein-ligand interaction score table, the energy values range between -1.45 kcal/mol and 2.12 kcal/mol (Table 9-4). These values correspond with the average interaction energy of a ligand (averaged over all ligand atoms and all atom types) with a particular class of atom triplet. Interestingly, the strongest interaction energy of -1.45 kcal/mol is very close to the maximum affinity value of -1.50 kcal/mol per ligand atom which was estimated from an analysis of experimental binding data [218]. Similarly, these energy values ranged between -2.22 kcal/mol and 1.80 kcal/mol for the protein – protein interaction dataset (Table 9-2); and -1.88 kcal/mol and 1.13 kcal/mol for the protein – peptide interaction datatest (Table 9-3).

5 Applying Computational Methods to Blys/BAFF interaction

5.1 Introduction

5.1.1 The TNF superfamily

The superfamily of tumor necrosis factor cytokines (TNF) and their corresponding receptors (TNFR) constitutes a class of cell-signaling molecules which regulate essential biological functions such as cell proliferation, survival, differentiation, tissue homeostasis and apoptosis. The majority of these TNFs are predominantly expressed by immune cells. The TNF/TNFR superfamily (Figure 5-1) constitutes 19 cytokines and 29 receptors [219]. Many of the TNF cytokines exert their functions either as type II transmembrane proteins and/or in soluble form by binding to one or more of the TNF receptors. The TNFRs are type I transmembrane proteins which often also exist as soluble proteins [220].

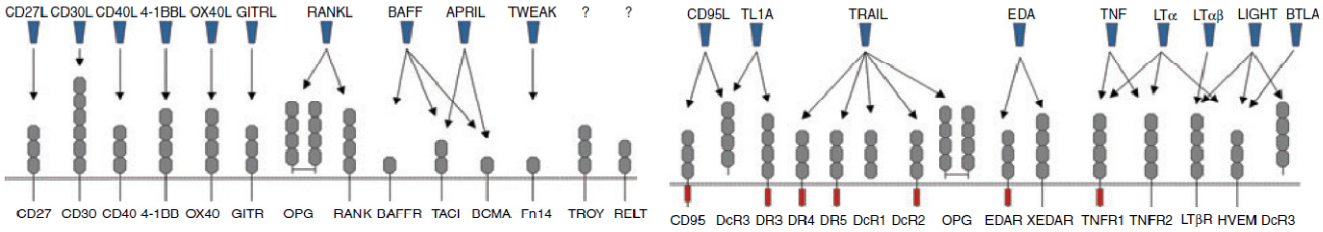

Figure 5-1: The TNF/TNFR superfamily (blue: TNFs, grey: TNFRs with the appropriate number of CRDs, red: death domains [220].

5.1.2 Structural and functional characteristics

The active form of TNF-superfamily cytokines is as self-associating trimers. They share a relatively low primary sequence identity (20% to 30%) leading to diverse ligand/receptor interfaces. These diverse interfaces are responsible for the observed receptor/ligand specificities [221-223]. Avariety of TNF-family orthologues for BAFF, *April*, *Tweak* (Figure 5-1) etc. in zebra fish and other teleosts have been discovered through phylogenetic analyses [224].

The TNF receptors are characterized by the presence of one or more extracellular Cys-rich domains (CRDs). The presence of the multiple disulfide bridges in the extracellular parts of the TNF receptors leads to the formation of relatively rigid, highly constrained loops, responsible for the interaction with the TNF-superfamily cytokines. TNFRs (Figure 5-1) are classified into three different classes, according to the presence of either an intracellular death domain (*Fasr, Trail-R1* ...)[225] , or TNF Receptor Associated Factor (TRAF) interacting motifs, TIMs (*BaffR*, *Rank*...) [226] or no functional intracellular signaling domain at all (*OPG, DcR1*) [227].

5.1.3 TNF cytokines and disease

A few members of the TNF superfamily have been implicated in the development and progression of multiple diseases in the fields of autoimmunity, neurodegenerative diseases, bone destruction, liver diseases, and cancer [228]. The B lymphocyte stimulator (*Blys*) is a key survival factor for B lymphocytes [229]. Mice having their *Blys* knocked out lack mature B cells in peripheral lymphoid tissues [230], overexpression of *Blys* in transgenic mice produces key-symptoms of

autoimmune diseases and *Blys* plasma levels in patients with several autoimmune diseases [231] have been reported to correlate with disease burden [232].

A proliferation inducing ligand (*April*) shares approximately 30% sequence identity to *Blys* in the TNF domain and is closely related to *Blys*; *it* recognizes two of the three reported *Blys* binding TNF receptors, *Taci* and *Bcma*. *Blys* (Figure 5-1) and has been implicated in autoimmune diseases such as Multiple Sclerosis, Systemic Lupus Erythematosus, Sjogren's Syndrome and Rheumatoid Arthritis [233, 234]. Furthermore, *April* was found to promote tumor cell survival *in vitro* and in tumor transplant *in-vivo* models. Both *Blys* and *April* have been found to play crucial roles in hematological malignancies [235]. Other TNFRs and their associated diseases include: *Tweak* (chronic immune diseases, arterioscleroses, and cancer), *Rank* (bone metastases and multiple myeloma) and LIGHT (intestinal inflammation and arthritis) [220, 236, 237].

In addition to their clear correlation with disease, TNF superfamily members are promising drug targets. Several TNF-receptors (*Taci*, *Bcma*, and *BaffR*) consist of only one or a partial Cys-rich domain. The extracellular parts of these receptors consist of highly constrained, short peptide stretches. Therefore, these receptors represent ideal examples of rigid natural ligands for docking and virtual screening experiments. Furthermore, the protein/protein interaction surface of TNF-receptors with their TNF-cytokines members (*Blys*, *April* and *Tweak*) is exceptionally small but of high affinity (low nanomolar to sub-nanomolar). The best studied example is

Blys/*BaffR*, where a highly constrained 26 aminoacids core region of the receptor has been shown to be sufficient for high affinity (Kd 70 nM) binding [238]. The essential recognition motif comprises a 6-residue hot-spot structured as a β-hairpin loop. This characteristic loop motif of the interaction hot-spot is presumably shared at least by *April* and *Tweak* (Figure 5-2).

5.1.4 Target validation / medical need

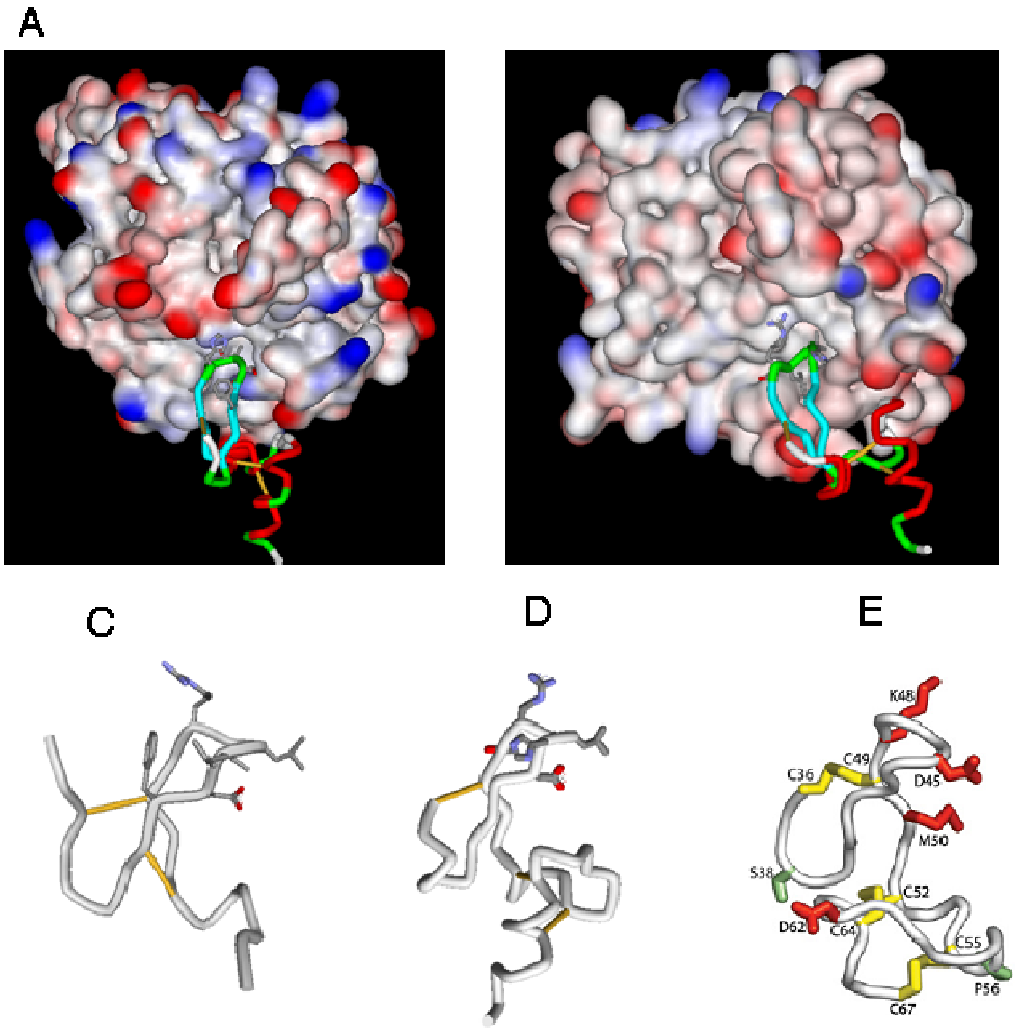

Figure 5-2: Binding modes of *Bcma*; *Taci* and FN14 to *Blys*, *April* and *Tweak*. A) (pdb: 1OQD) shows the binding of *Bcma* (sticks) to *Blys* (surface). B) (pdb: 1XU1) shows the binding of *Taci* (sticks) to April (surface). C) (pdb: 1P0T) shows the structure of *BaffR*. D) (pdb: 1XU1) shows the structure of *Taci*. Finally, D) shows the homology model for Fn14 and while binding to *Tweak* [239].

Current strategies for targeting TNF-superfamily members focus on the development of biologics i.e. neutralizing or agonistic antibodies. Several TNF-targeting products have already been marketed: Remicade®, Enbrel®, and Humira®. Several biologics for targeting *Blys* and/or *April* are currently in development: Belimumab (a Human anti-*Blys* monoclonal antibody, GlaxoSmithKline, Human Genome Sciences, Phase III), *Taci*-Immunoglobin (*Taci*-Ig) (soluble *Taci*-Ig, ZymoGenetics/Serono, Phase II), AMG623 (Amgen, *Blys* targeting peptide-fusion protein, Phase I), BR3-Fc (Genentech, Soluble BAFF-R-Immunoglobin, Phase I). However, small-molecule ligands/peptides would provide a significant economic, scale, and bioavailability advantage over antibodies and whole-protein treatment. High affinity peptides (larger than 12-mers), derived from phage display studies have been described for *Blys* [240]. Furthermore, moderate affinity TNF-α trimerization inhibitors and inhibitors [241] of the TNF-α intracellular signaling cascade have also been described [242].

5.1.5 The Integrated Chemical Biophysics (ICB) Process

Current experience in the protein-protein interactions field has shown that high throughput screening (HTS) does not routinely identify compounds that disrupt protein interactions [243]. However several starting points were identified from HTS with large compound collections (>250 000) to identify moderate hits (Ki in the mid micromolar range). Reasons for the limited success rate of routine HTS are:

a) The large interaction surfaces 1,500-3,000 $Å^2$ (compared with those of small molecule-protein interactions of 300-1000 $Å^2$)

b) limited structural starting points (protein-protein interactions have proven to be highly adaptive and hence the best binding sites and modes can not

be often observed from static structures of either free protein target or protein-protein complexes)

c) Current compound collections are extensively biased towards ATP mimics and GPCR binders [244-248]. Recent experience suggests that compound collections with higher molecular mass and significantly different design are required for tackling protein-protein interactions [243, 249].

As protein-protein interactions are highly diverse, with only the most related proteins sharing common features, it has been proposed that target biased libraries will be needed for each interaction in order to derive optimized inhibitors.

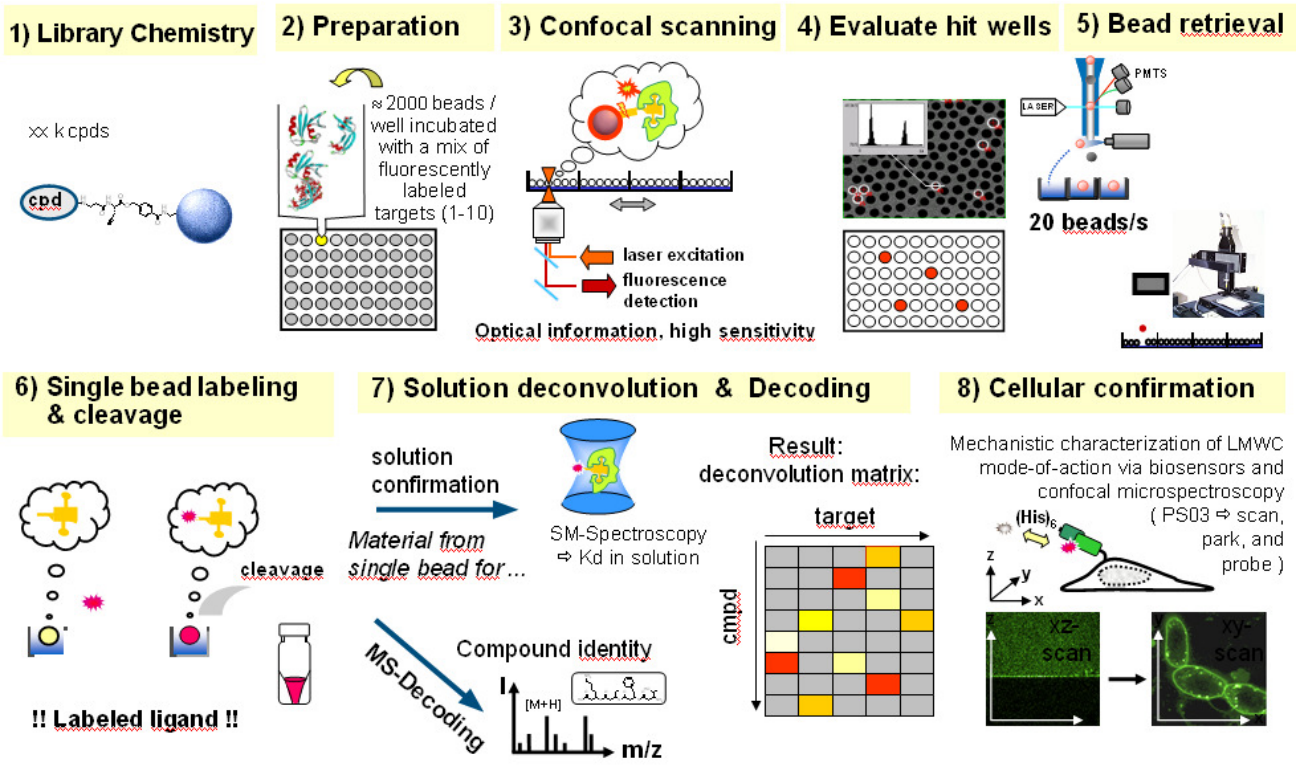

Figure 5-3: The ICB-process

The ICB-process (on-bead library synthesis, on-bead screening, solution conformation, and cellular validation (Figure 5-3)) as developed by Auer et al. will provide an ideal technology starting point for targeting protein-protein interactions such as TNF-superfamily members due to:

1. The possibility to generate project tailored libraries by on-bead synthesis of peptides, cyclic peptides and peptidomimetics with non-natural side chains and building blocks
2. The ability to obtain ligand series with proven Kd's and initial Structure – Activity Relationship (SAR) experiments from each primary screening round with minimal (ca. 50 picomols) substance without extensive re-synthesis.
3. The ability to generate fluorescently labeled binders in combination with high-resolution imaging and micro-spectroscopy for resolving details of the signaling mechanisms and specificities of the TNF/TNFR superfamily.

Previous experiments using *Blys* as a target have yielded mid-micromolar affinity (Kd) compounds based on a β-peptides with β-turn foldamers. These compounds are the first low molecular weight ligand binding inhibitors for one of the TNF-superfamily members (a minimal binding motif for *Blys* has previously been reported to contain 12 amino acids, however its high-affinity binding was not reproduced in our hands).

5.2 Strategy

This work focuses on the computational methods focused at designing inhibitors for the *Blys:BaffR* interaction. The project outline, the role of computational biology, the

specifics of the *Blys:BaffR* interaction site, and the proposed starting models for designing inhibitors are discussed below.

5.2.1 General Project Flow

The proposed project flow consists of three main project tracks:

1. Design of new on-bead turn-mimicking peptidomimetic libraries, virtual hypothesis testing by docking experiments, prioritization of proposed library designs, synthesis and on-bead screening of selected libraries.
2. Virtual screening of existing virtual compound libraries, synthesis of a small compound collection around best hits on-bead for experimental testing.
3. Partnering with industry for competition screening of existing compound archives to obtain low-to-medium affinity starting points for further optimization by bead based synthesis and screening cycles.

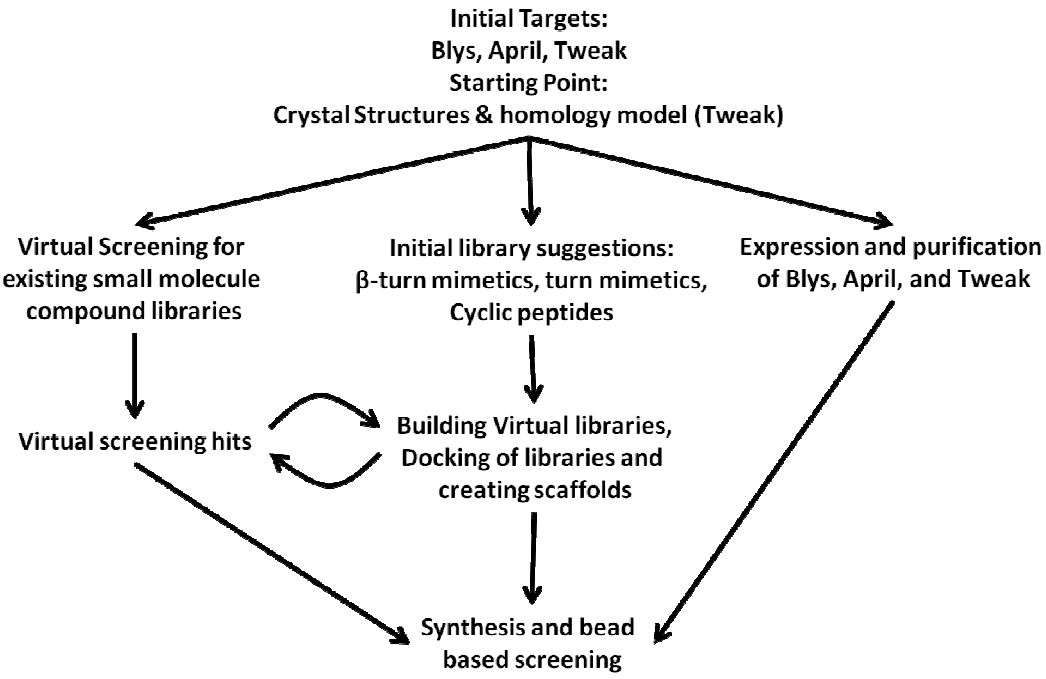

Figure 5-4: General Project Flow

Following this plan, this project is expected to produce novel TNF/TNF-ligand crystal structures and provide valuable insight into targeting of other TNF family members.

5.2.2 The Role of Computational Biology

This work deals with the computational biology aspect of this project. The target will be the generation of several compound libraries starting from different starting models. The proposed methodology to reach this goal is:

1. Search for possible inhibitor compounds using the program LIDAEUS [69]
2. Search for possible inhibitor compounds using the programs UFSRAT (Steven R. Shave) and Autodock [71].
3. Design Peptides to inhibit the interaction
    a. Model choice/optimization
    b. Mutagenesis of side chains of the binding face to test Model viability
    c. Energy Minimization
    d. Scoring Fits and models and Deciding on the best Models to be used
    e. Creating the library

5.2.3 Study of the Blys Binding Site

The original binding site is located in PDB structure 1OSG, with chains 'D' and 'E' on the protein side of the molecule while chain J resembles the original peptide that docks in the groove (Figure 5-5). The peptide in chain J is immobilized on the cell surface. The protein recognizes that peptide and docks on the cell surface.

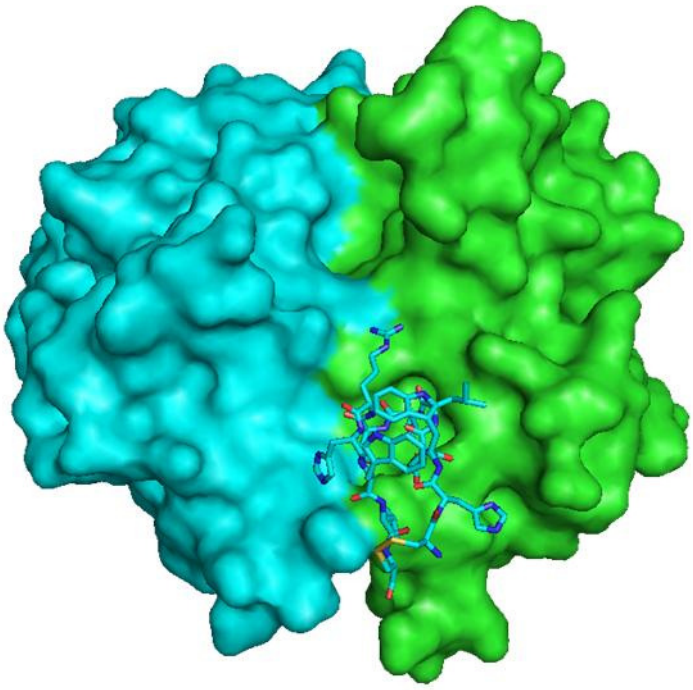

Figure 5-5: The Binding Site of the protein 1OSG. Figure shows Chains D and E (Green and Blue) with the peptide (chain J, here in sticks) bound to the binding pocket.

Study of the interaction between the protein and the peptide shows that the most important peptide residues in this interaction are the DLLVR (residues 26 – 30), with the L-27 not playing an equally important role in the interaction. These interactions are summarized in Table 5-1.

Table 5-1: Important interactions in the *Blys*/*BaffR* complex

Peptide (*BaffR*)	Protein (*Blys*)
Asp 26	Arg 265 of Chain D
Leu 28	Met 208 and Gly 209 of Chain D
Val 29	Asn 242 (<4 Å) of Chain E Ala 207 (4.53 Å) of Chain D (to be optimized)
Arg 30	Asp 257 (2.6 Å) of Chain E

In order to mimic and eventually inhibit this interaction, a search for possible inhibitors was conducted. As a general guideline, it is important for any designed/discovered candidate inhibitor to mimic the interactions discussed above. Moreover, the Leu residue is inside a cleft on the protein surface, and a possible enhancement to this interaction is to have this residue going deeper into the groove (Figure 5-6). Those guidelines were used to filter out the virtually designed ligands and find out the best sequences and structures. This is discussed in Section 5.4.

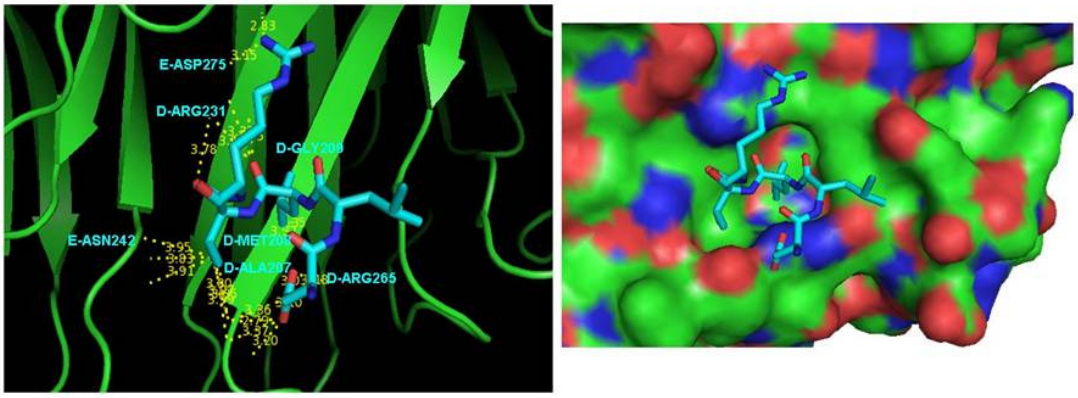

Figure 5-6: Important interactions in the *Blys*/*BaffR* complex. Mimicking these interactions is important for inhibiting the function of the complex. Left image shows the cartoon representation of Blys and the sticks representation of BaffR. Distances between Blys atoms and BaffR atoms within 4Å are labeled in yellow and the residues on the Blys side are marked (D-ASP 275 is ASP 275 of chain D). The figure on the right shows the surface of Blys and how the BaffR reaches deep within the pocket on the surface.

5.2.4 Starting Models for Peptide Design

Several starting models have been considered to use as scaffolds where different side chains could be installed. The designed peptides should exhibit a β-turn, mimicking the original ligand. On this assumption, a ligand library was constructed based on a starting model of cyclic hexapeptides.

Cyclic hexapeptides

Due to the cyclic nature of those peptides, they should exhibit a similar turn to that of the β peptides. The only difference will be that the turn comprises 8 members rather than 10. However; cyclic α peptides are still good candidates for mimicking the *Blys*-BAFF interaction.

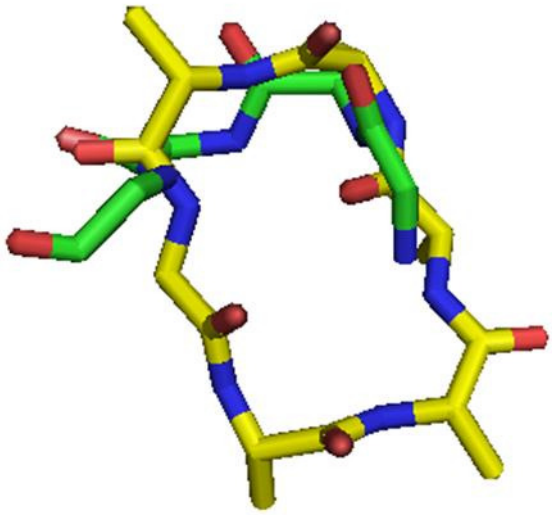

Figure 5-7: Backbone alignment of a cyclic α peptide (yellow) with the original ligand (green) shows that the β-turn can be mimicked with a cyclic hexapeptides.

α peptides have the advantage of being well researched, and therefore the side chain torsion angles can be more accurately predicted by the algorithms used. However, α peptides pose the challenge of finding a stable sequence that is not hydrolyzed quickly by the body. This is vital if any of the designed ligands are to be used as a drug. The use of irregular side chains might make it harder for the body enzymes to recognize the molecules.

5.3 Searching for possible inhibitor compounds using the program LIDAEUS

One strategy to tackle inhibiting the *Blys*/*BaffR* interaction is to check for chemical compounds that are likely to bind to the binding site in *Blys*. LIDAEUS is the best tool to conduct virtual screening experiments of this kind. In coordination with EDULISS, LIDAEUS has access to around 4 million compounds that are available in several supplier catalogues. A set of site points was generated based on the original ligand interaction with *Blys*, and is shown in Figure 5-8. Then LIDAEUS was run on the *Bluegene* supercomputer at Edinburgh University and set to return the best 1000 hits.

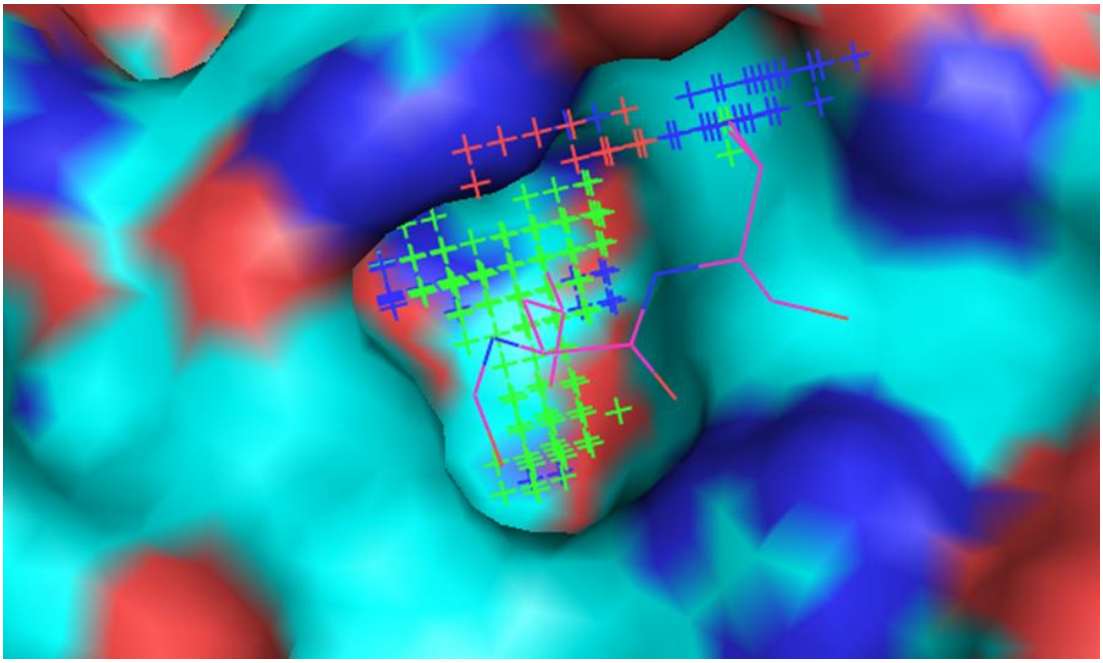

Figure 5-8: The site points generated by LIDAEUS for the *Blys*/*BaffR* complex. These Site Points indicate the most favorable type of atoms to be positioned at a certain location. Green indicates hydrophobic, Blue indicates Hydrogen Bond Donors, and Red indicates Hydrogen Bond Acceptors. The binding site of protein 1OSG (chains D and E) is shown as a surface and the bound peptide is shown as sticks.

The top 200 hits were analyzed in an attempt to select good drug candidates that would inhibit the *Blys*/*BaffR* interaction. Lideaus appends a rough estimate of the binding enthalpy to each selected compound (ΔH). The top 200 hits had a ΔH in the range of [-43.155 kcal/mol, -37.097 kcal/mol]. Values of -20 and below are usually considered good and worth checking. A few steps were taken to filter out molecules that are not likely to be good drug candidates. Sugars and large molecules were thus discarded. The remaining compounds were examined to decide whether they are worth buying or not. Three molecules were selected; they are referred to by their EDULISS IDs: 25SPH1-104-152, 42SPH1-002-100, and 28SPH1-443-207 (Figure 5-9). As denoted in Figure 5-9, those compounds will be referred to as compounds A, B, and C.

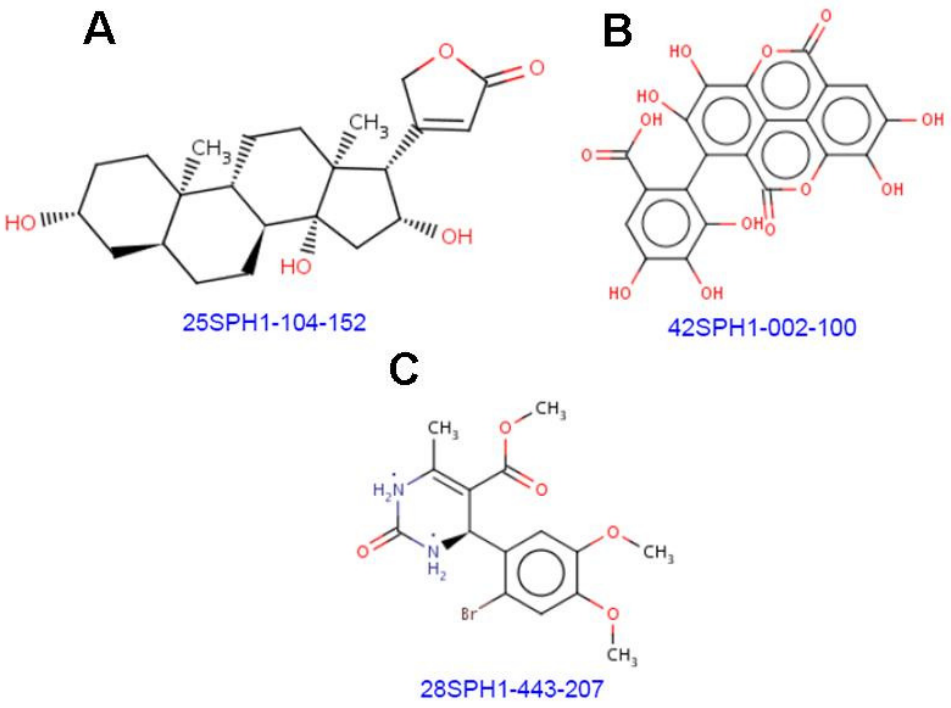

Figure 5-9: Interesting compounds from the LIDAEUS virtual screening experiments.

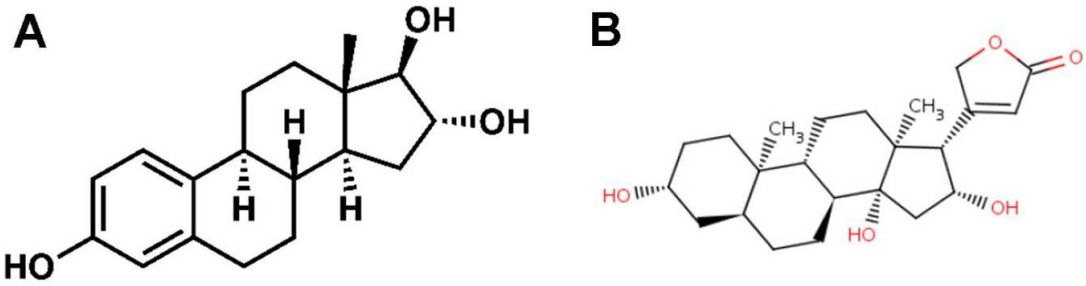

Figure 5-10: The similarity between Estriol (Figure A) and Compound A (Figure B). Estriol is one of the three main estrogens produced by the human body

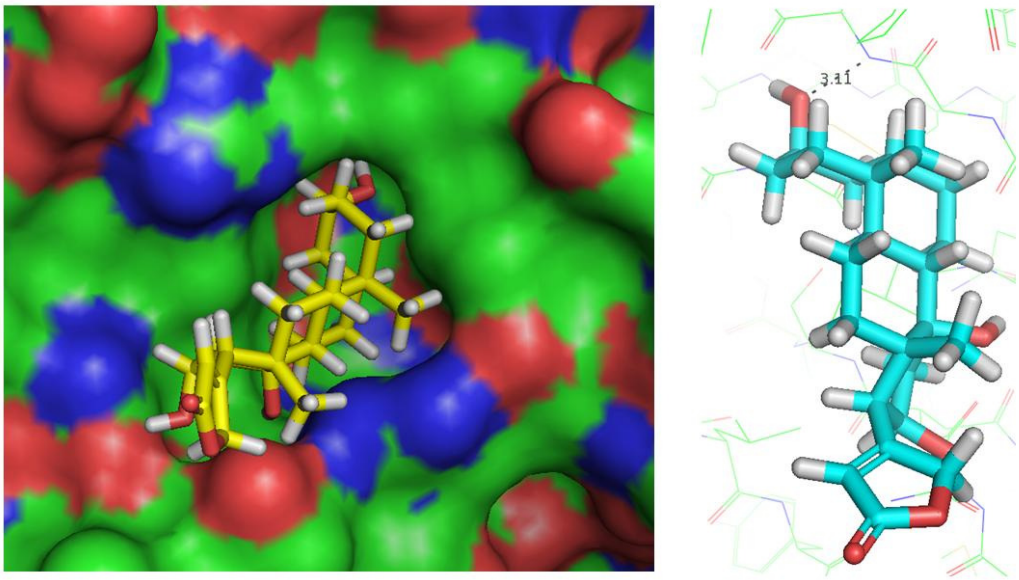

Figure 5-11: The predicted binding between Compound A (Gitoxigenin) and *Blys* (notice the geometrical compatibility). The figure on the right shows a possible hydrogen bond deep inside the binding pocket.

Compound A (named in the Sigma-Aldrich database as Gitoxigenin) is a steroid. Its strong resemblance to Estrogen (Figure 5-10) poses a lot of question marks about how useful it will be as a drug candidate. However, and as shown in Figure 5-11, this compound fits into the binding site of *Blys* in a geometrically perfect way. Its

calculated ΔH is -40.891 kcal/mol and can be cheaply acquired from Sigma-Aldrich (15.9 GBP for 10 mg). It has therefore been chosen for wet lab verification. Compound C (Figure 5-12) also shows a good and deep binding into the *Blys* pocket. It is found on Ambinter's website under the id of 6329476.

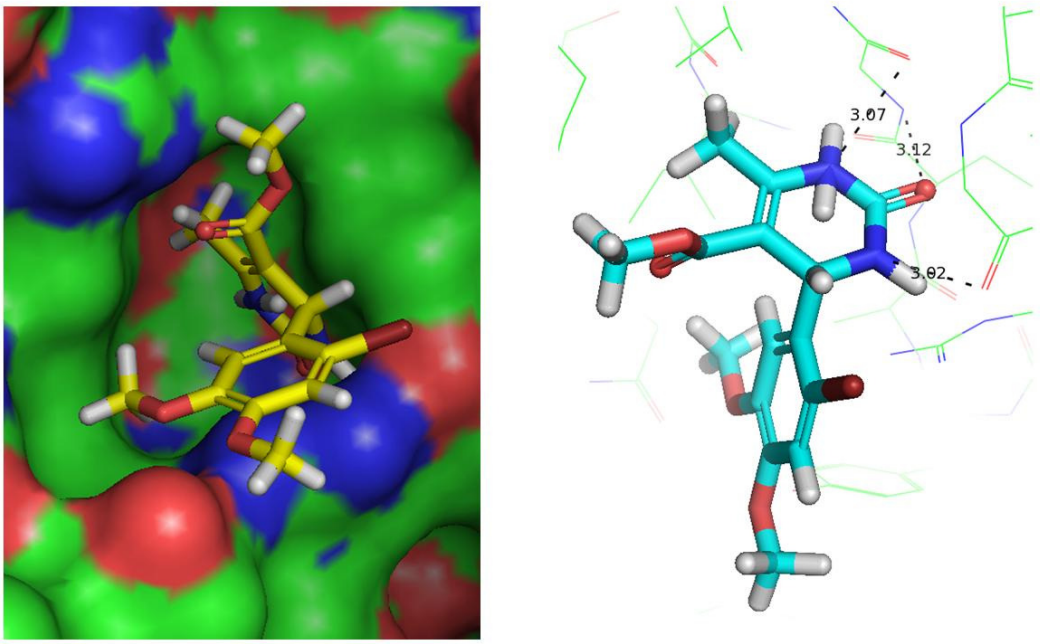

Figure 5-12: The predicted binding between Compound C and *Blys*. The figure on the right shows the possible hydrogen bonds between the protein and the ligand

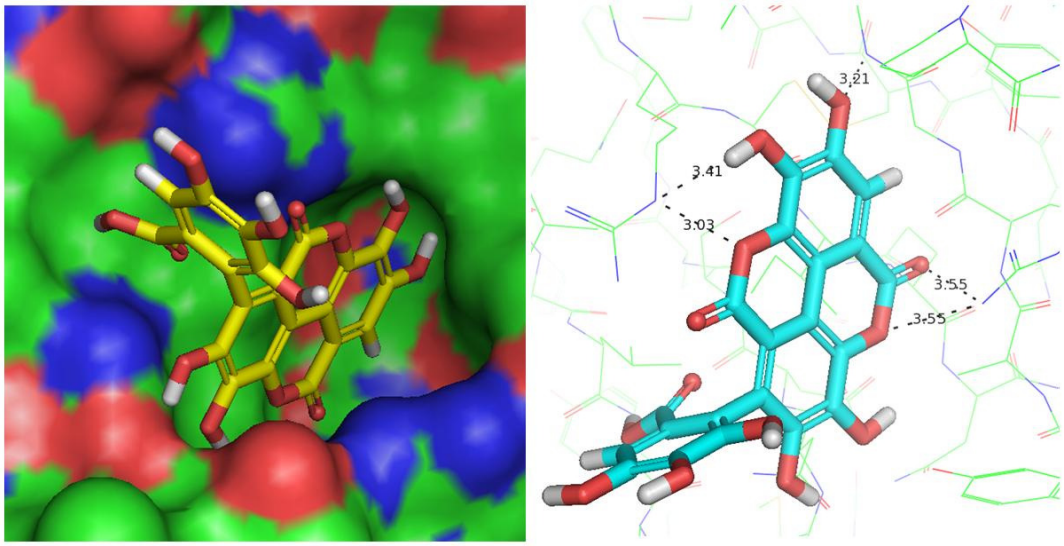

Figure 5-13: The predicted binding between Compound B and *Blys*. The figure on the right shows the possible hydrogen bonds between the protein and the ligand.

Compound B (named MEGxp on the Analyticon database) fits deeply into the pocket of *Blys* and has appeared 3 times in the top 205 compounds, each time being in a different orientation (Figure 5-13). These three orientations were scored at ΔH values of -38.51 kcal/mol, -37.136 kcal/mol, and -37.300 kcal/mol. This compound also has a unique attribute: the presence of 4 benzene rings attached to each other in a rhombus pattern. This attribute renders the compound to be easily distinguishable on the binding assay used (fluorescence anisotropy assay / polarization assay). Unfortunately, the supplier firm (Analyticon) has a 500 EUR minimum invoicing price, which will render this molecule too expensive although its real price is 31 EUR per 1 mg. Four benzene rings were then searched for in the EDULISS database using the program UFSRAT. When given the entire compound, UFSRAT did not find a suitable similarity with any of the compounds. The 4 benzene rings where then searched on their own, and UFSRAT found an exact match (EDULISS ID 25SPH1-

179-824), called Ellagic Acid in Sigma. Ellagic Acid was docked with LIDAEUS and gave an acceptable ΔH of -22.566 kcal/mol; presenting a good workaround for compound B. Compound B was also sent to Novartis to be checked against their private database.

Moreover, the LIDAEUS search showed that a singly or doubly OH substituted benzene ring is a good binding motif for the *Blys* pocket. This motif could be a good basis for designing side chains for the peptide library (favoring a Tyr residue and its non-natural side chain analogues as possible mutations for Leu 28).

5.4 Designing inhibitor Peptides

5.4.1 Mutation Program

The Mutation program is a vital component for creating the virtual database. It takes in a polypeptide in PDB format, and a text file that instructs it about which residues to mutate, the chi angles to use for the side chains, and the choice of substitute side chains to be used per location. The program will then fix the required side chains in the desired locations and output the resulting structures in Sybyl Mol2 format. The program will output 1 structure for each permutation. That means if we have n residues positions to be mutated, each with m different side chains in mind, with a specification of using r Chi-Angle rotamers for the Chi-1 angle (the dihedral angle between the 4 points C, N, α C, and β C), the result will be $(m \times r)^n$ different structures.

5.4.1.1 Pseudocode

The Program is of a recursive nature. Recursion was a necessary choice to be able to cope with a variable number of positions to be mutated. The Pseudocode of the algorithm used to generate the molecules is given below.

Algorithm 5-1: The Mutation Program that constructs different molecules based around a specific cyclic hexapeptide backbone with a list of side chain choices at each residue position.

```
Mutate(p, n)
        IF (p > n)
                Return
        FOR each substitute side chain choice s:
                Fix s in the position of p, overlapping N, C_α, and C_β atoms.
                Get the rotamers from the database
                FOR each rotamer choice r:
                        Adjust side chain atoms according to r
                        IF p is the position to be mutated
                                Print Molecule
                        ELSE
                                Mutate (p+1, n)
                        END_IF
                END_FOR
        END_FOR
END_Mutate
```

5.4.1.2 Geometrical Operations in the Mutation Program

The operation of substituting the side chains at a certain position with other suitable ones in order to mutate the residue is done in the following way:

1. Read in a template molecule with the desired side chain
2. Delete the side chain atoms of the residue to be mutated, except the β-carbon
3. Generate a transformation matrix that would transform the atoms N, C_α, and C_β from the substitute aminoacid onto the atoms N, C_α, and C_β of the target aminoacid

4. Apply the transformation for each side chain atom in the substitute amino acid to get the final coordinates

The geometric operations needed to find the transformation matrix that would transform the atoms N, C_α, and C_β of the substitute aminoacid onto those of the target aminoacid are:

1. Let M1 be the transformation matrix that puts N, C_α, and C_β of the target aminoacid residue in the XZ plane with C_α on the origin and C_β on the Z-axis
2. Let M2 be the transformation matrix that puts N, C_α, and C_β of the substitute aminoacid residue in the XZ plane with C_α on the origin and C_β on the Z-axis
3. To transform the atoms of the substitute residue onto that of the target residue, apply transformation matrix M = inverse(M1) × M2.

The geometric operations needed to find the transformation matrix that puts N, C_α, and C_β of an aminoacid residue in the XZ plane with C_α on the origin and C_β on the Z-axis are:

1. Let M1 be the translation matrix that puts the α-carbon at the origin.
    a. Also let
        i. $C_{\alpha x}$ be the x coordinate of C_α
        ii. $C_{\alpha y}$ be the y coordinate of C_α
        iii. $C_{\alpha z}$ be the z coordinate of C_α

$$M1 = \begin{bmatrix} 1 & 0 & 0 & C_{\alpha x} \\ 0 & 1 & 0 & C_{\alpha y} \\ 0 & 0 & 1 & C_{\alpha z} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

b. Apply M1 to N, C_α, and C_β

c. Normalize vector ($C_\alpha C_\beta$)

2. Let M2 be the transformation matrix that puts the β-carbon on the Z-axis

   a. Also let

      i. $C_{\beta x}$ be the x coordinate of C_β

      ii. $C_{\beta y}$ be the y coordinate of C_β

      iii. $C_{\beta z}$ be the z coordinate of C_β

      iv. $d = \sqrt{(C_{\beta x}^2 + C_{\beta y}^2)}$

$$M2 = \begin{bmatrix} C_{\beta z} & 0 & -d & 0 \\ 0 & 1 & 0 & 0 \\ d & 0 & C_{\beta z} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} C_{\beta x} \div d & C_{\beta y} \div d & -d & 0 \\ -C_{\beta y} \div d & C_{\beta x} \div d & 0 & 0 \\ d & 0 & C_{\beta z} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The matrix on the right rotates C_β around the Z-axis, putting it in the XZ plane (setting y=0). The matrix on the left rotates C_β around the Y axis, putting it in the YZ plane (setting x=0). The combination of those 2 matrices will put C_β on the (XZ ∩ YZ) plane (which is the Z-axis).

b. Apply M2 to N, C_α, and C_β

c. Normalize vector ($C_\alpha N$)

3. Let M3 be the transformation matrix that puts N in the XZ plane

   a. Also let

      i. N_x be the x coordinate of N

      ii. N_y be the y coordinate of N

$$M3 = \begin{bmatrix} N_x & N_y & 0 & 0 \\ -N_y & N_x & 0 & 0 \\ d & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

4. The resulting matrix would be: M = M3 × M2 × M1

5.4.2 Energy Minimization

After the generation of the various permutations and conformations of the molecule, a fast energy minimization procedure is necessary to make sure the final models are reasonable and depict the reality in a correct and accurate manner. The Minimax program (part of the witnotp suite [250]) was used to perform this task. Minimax utilizes the Broyden-Fletcher-Goldfarb-Shanno (VA13A, Harwell) algorithm to perform molecular mechanics simulations, making the designed molecules more stable and hence the design more accurate.

5.4.3 Scoring Fits and Models

The best way to assess the quality of the starting model is by scoring how well those models resemble the original ligand. After each starting model has been permutated with a few side chains and minimized, the backbone of each permutation was taken in addition to the β-carbons, and a best fit was calculated. This best fit then reports a root mean square deviation (rmsd) value that signals how close this backbone is to the original ligand. Details of how the permutations where created for each starting model are found in later sections.

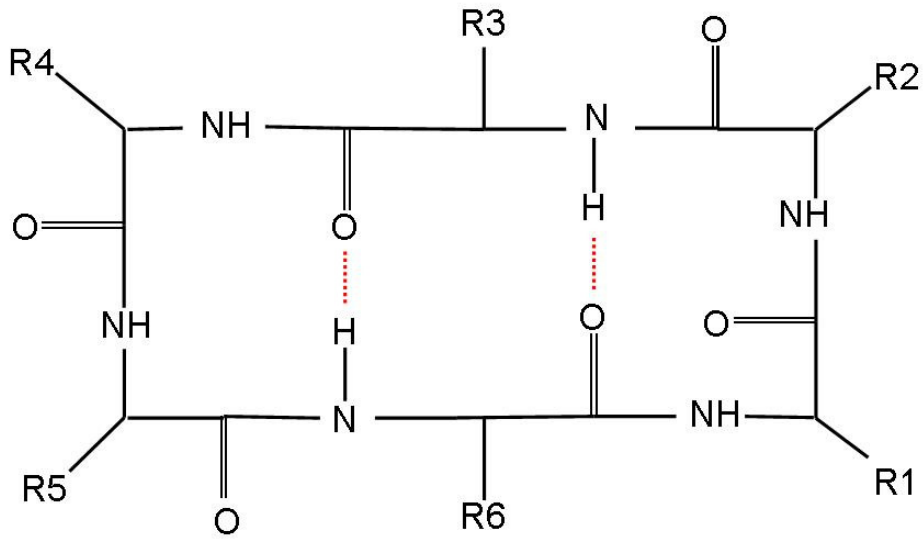

Figure 5-14: The basic structure of a cyclic hexapeptide. The dotted lines represent the backbone hydrogen bonds stabilizing the structure. The bead is to be connected at either R1 or R2.

5.4.4 Cyclic Alpha Peptides

Cyclic hexapeptides are good candidates for mimicking the β-turn motif of the *BaffR*. They are easier to synthesize and have a longer duration of activity compared to linear peptides [251]. Such structures exhibit 2 β-turn motifs and are stabilized by a pair of hydrogen bonds (Figure 5-14). One of those β-turn motifs will mimic the binding face of the *BaffR* (4 residues from the hexapeptide). A fifth residue will be fixed as an Ala in order to connect it to the bead. The sixth residue will be free and is dependent on the backbone from which the virtual library will be constructed.

The residues of any cyclic hexapeptides will be referenced as per Figure 5-16. R3, R4, R5, and R6 of the hexapeptide will constitute the binding face that will mimic Residues 26, 28, 29, and 30 of *BaffR*. Whilst those residues are to be mutated and the entire structure minimized to check if we have a good fit with *BaffR*, it is essential that residues 3 to 6 are not originally Gly or Pro. Gly exhibits backbone flexibility due to the absence of a side chain, making the original model too flexible to use for modeling (especially in the binding face residues). Having Pro in the binding face of a hexapeptide model is also unfavorable as Pros exhibit backbone rigidity due to the side chain coupling with the nitrogen atom. This rigidity will not be mimicked by the other aminoacids. With one of the remaining two positions (R1 and R2) restricted to Ala (for bead connection), that leaves only one free residue that could virtually be any aminoacid. That means that while searching for hexapeptide models, only one Pro or Gly residue should be present in the original sequence of the model.

Table 5-2: Cyclic hexapeptides in the Cambridge Crystallographic Database, their sequences and number of Pro/Gly residues.

ID	Sequence	Count(Pro, Gly)
AAGAGG10	Ala-Ala-Gly-Ala-Gly-Gly	3
BAMLIK	Gly-His-Gly-Ala-Tyr-Gly	2
BIHTUH	Phe-Pro-Ala-Phe-Pro-Ala	2
BIHXUL10	Pro-Val-Phe-Phe-Ala-Gly	2
BINJIR	Gly-Pro-Pro-Gly-Pro-Pro	6
BUNYEO10	Pro-Pro-Gly-Pro-Leu-Gly	5
BUYXOI	Pro-Pro-Pro-Pro-Pro-Pro	6
CAHWEN	Phe-Leu-Gly-Phe-Leu-Gly	2
CAMVES	Pro-Pro-Ala-Ala-Ala-Ala	2
CGDLLL10	Gly-Leu-Leu-Gly-Leu-Leu	2
CGLEGL	Gly-Leu-GlyE-Gly-Leu-Gly	3
CGLPGL	Gly-Pro-Gly-Gly-Pro-Gly	6
CLPGDH	Leu-Phe-Gly-Leu-Phe-Gly	2
CYBGPP	Gly-Pro-Phe-Gly-Pro-Phe	4
CYHEXG	Gly-Gly-Gly-Gly-Gly-Gly	6
DUPKEE	Phe-Pro-Ala-Phe-Pro-Ala	2
DUYTIA	Pro-Phe-Phe-Pro-Phe-Phe	2
GAJFAY	Pro-Phe-Thr-Phe-Trp-Phe	1
GGAAGG	Gly-Gly-Gly-Gly-Ala-Ala	4
KIVDIC	Val-Pro-Ala-Val-Pro-Ala	2
PAPRVA	Phe-Pro-Val-Phe-Pro-Val	2
ProGly20	Pro-Gly-Pro-Gly-Pro-Gly	6
RAQVOU	Gly-AIB-Gly-Gly-AIB-Gly	4
YEXJIV	Ala-Pro-Gly-Phe-Val-Ser	2
ZAJPAB	Ala-Gly-Val-Pro-Val-Trp	2
ZUKRAY	Gly-Thr-Phe-Leu-Tyr-Val	1

Two searches were conducted to find the proper hexapeptide models. The first was in the protein data bank, and returned zero hits. There were 4 cyclic hexapeptide structures in the PDB (1SKI, 1SKL, 1SKK, 1QVK), but they were NMR models rather than X-ray structures and were neglected. The second search was conducted in the Chemical Database Service (CDS, part of the Cambridge crystallographic database) and returned 26 X-ray structures (Table 5-2). Those structures were filtered by the number of Pro or Gly residues they have and limited to those with only one (based on the discussion above). This gave two structures to check as possible

templates for the peptide design experiment to construct the virtual library: GAJFAY and ZUKRAY. Figure 5-15 shows those 2 structures superposed over the original ligand.

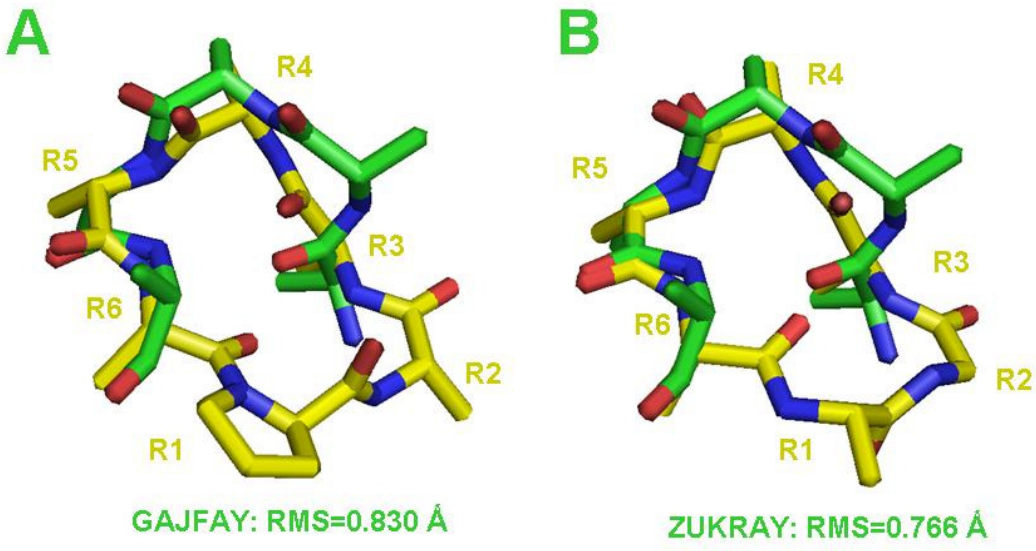

Figure 5-15: The superposition of the backbones of GAJFAY (A) and ZUKRAY (B) on the *BaffR* backbone. The rmsd values shown correspond to the superposition by Cα and Cβ atoms of R3, R4, and R5 and Cα atom of R6. In the case of R6, the β-carbon was neglected because of the difference in orientations between the models and the original ligand, which will be discussed in Figure 5-16.

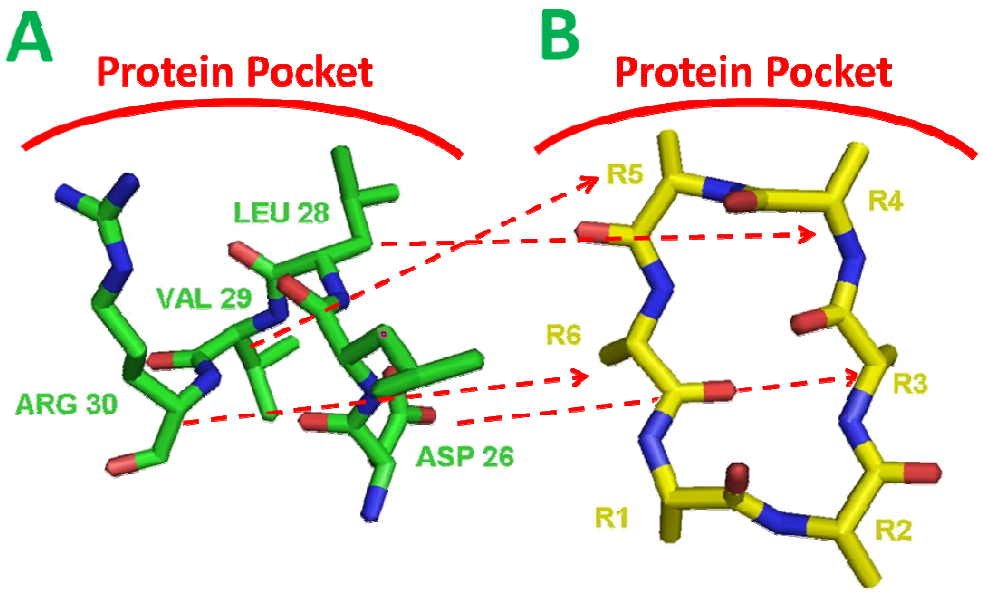

Figure 5-16: The difference in side chain orientation in Arg 30 of A and R6 of B. By changing the conformation of R6 into a D-amino acid conformation, the backbone in Figure B (representing both ZUKRAY and GAJFAY) will be more capable of mimicking that of *Blys* (Figure A). The arrows connect the BaffR amino acids and their corresponding positions on the cyclic hexapeptides.

All starting models were stripped of their side chains except the β-carbons at positions 3, 4, 5, and 6. The Pro at R1 in GAJFAY was kept as is and R2 changed into an Ala where the bead is to be connected. Similarly, the Gly at R2 in ZUKRAY was kept as is and R1 changed into an Ala where the bead is to be attached. As denoted in Figure 5-16, the binding face of the peptides is formed of 4 consecutive aminoacids. In the case of R5 which is supposed to mimic Arg 30 of the original ligand, the positioning of $C_\alpha C_\beta$ is not optimal as it's pointing sideways in both GAJFAY and ZUKRAY while it is pointing upwards in the original ligand. A change of stereochemistry is studied, switching R6 into a D-amino acid. GAJFAYD6 and ZUKRAYD6 are the homologues of GAJFAY and ZUKRAY, having a D-amino

acid at R6. Table 5-3 gives the original residue numbers in these structures referred to by R1 through R6.

Table 5-3: Distinct starting Models for Cyclic hexapeptides. The second column denotes the original residue numbers in the structures to position as R1, R2, R3, and R4.

Model	R1, R2, R3, R4
GAJFAY	3, 4, 5, 6
ZUKRAY	2, 3, 4, 5

To check which of these starting models is capable of mimicking the conformation of the original ligand, the binding face residues of these ligands were mutated to all aminoacids {Arg, Asp, Glu, Leu, Lys, and Val}. The remaining 2 residues were fixed as discussed above. Based on this permutation, each starting model will generate 6^4 = 1296 possible structures. Those structures were then minimized by Minimax to ensure they are in the most favorable energy conformations. A study on the strength of the backbone was conducted by calculating the rmsd between the minimized structures with their unminimized analogues. The rmsd was calculated as the mean rmsd of all the backbone atoms of the binding face residues in addition to the C_β atoms of these residues as well (Figure 5-17).

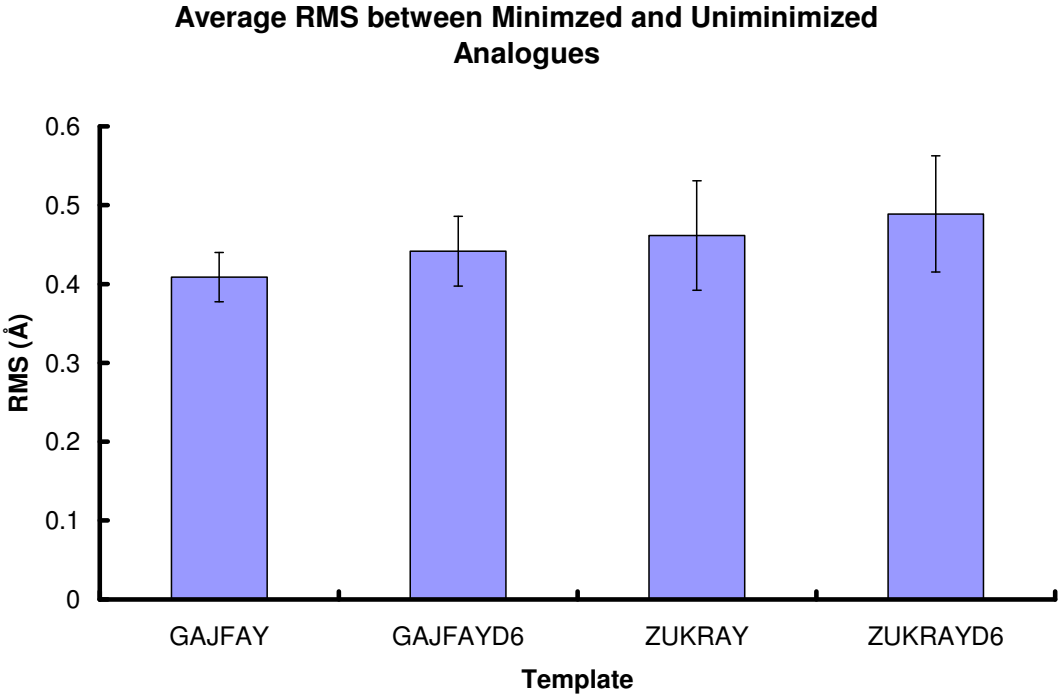

Figure 5-17: Each model produced 1296 different analogues, each having a unique sequence at the binding face of the molecule. Each of these analogues was minimized and the rmsd between the minimized and unminimized structures recorded. The values shown here correspond to the mean and standard deviation of the rmsd values recorded, per Backbone. GAJFAYD6 is GAJFAY with R6 in D-amino acid conformation.

The choice of the most suitable starting models was based on which model has a binding face with a backbone fairly close to the original ligand after being minimized and mutated with different side chains. For each of the 1296 permutations, an rmsd was calculated between the permutation and the original ligand. The rmsd values were generated based on the α and β-carbon atoms of the binding face residues (R3, R4, R5, and R6). Results are shown in the Figure 5-18.

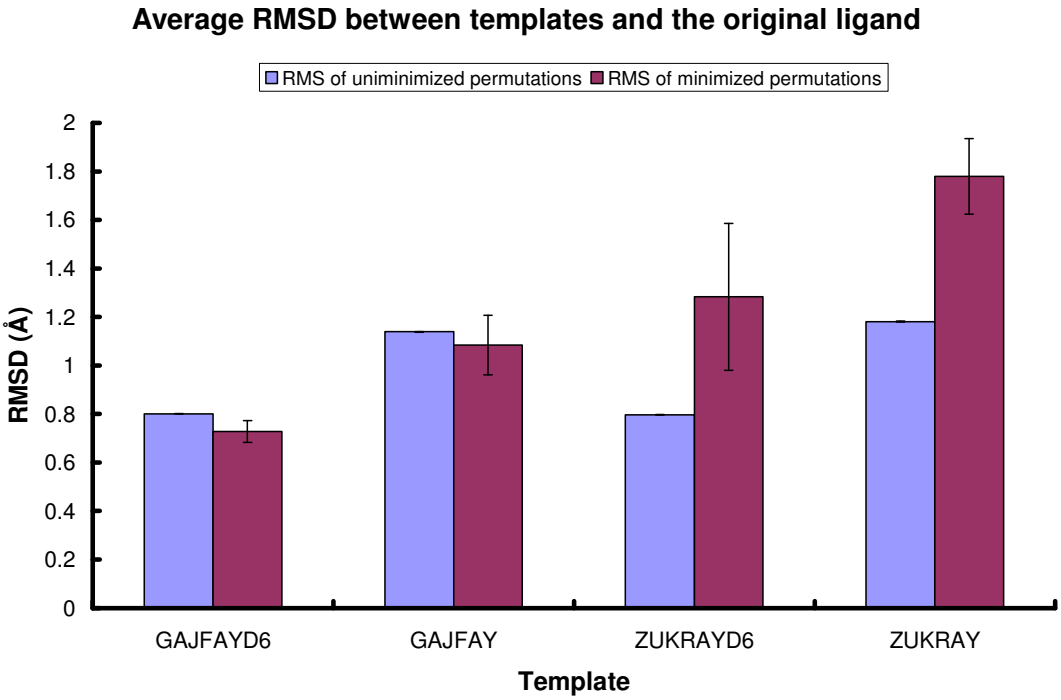

Figure 5-18: Each of the 1296 structures produced per model was superposed onto the *BaffR* using the α and β-carbons of R3, R4, R5, and R6. This figure shows the mean and standard deviations for the 1296 rmsd values recorded per model.

The studies done on the starting models showed that indeed, residues R6 should be in a D conformation. This is shown by both templates getting better positioning when in a D6 conformation. Although ZUKRAY initially had a lower rmsd than GAJFAY, the study showed that GAJFAY has a more stable backbone (Figure 5-17 shows that the rmsd between the minimized and unminimized analogues is lower in the cases of GAJFAY and GAJFAYD6 compared to ZUKRAY and ZUKRAYD6). This is probably due to GAJFAY having a Pro residues while ZUKRAY having a Gly residue in their sequences, the first giving the backbone more rigidity while the second making it more flexible. Even more, Figure 5-18 clearly shows that after studying the structures' behavior under several sequences (1296 to be exact), the minimized structures of both GAJFAY and GAJFAYD6 had lower rmsd values than

ZUKRAY and ZUKRAYD6.Consequently, all results point towards GAJFAYD6 as the optimal model to be chosen for building the cyclic hexapeptide library.

5.4.5 Building the library

Now that the model hexapeptide backbone has been chosen, the next step would be to optimize the side chain choices on this hexapeptide. To recap, the side chains at R1 and R2 are kept unaltered (as per GAJFAY). R3, R4, and R5 are used to mimic the original interactions contributed by Asp 26, Leu 27, and Leu 28 of the original ligand respectively. Two choices would then exist for R6: 1) as a D-Amino acid to mimic Arg 30 of the original ligand, and 2) as an L-aminoacid to mimic Val 29 of the original ligand. It was decided to focus on the D-amino acid choice as a start and then to explore the L-side chain choices later.

The choice of side chains was not restricted to the 20 natural aminoacids, but extended to several non-natural side chains. Other than increase the space of combinations we can produce, this also helps in further optimization of the interaction. For example, although the binding of Leu 28 in the hydrophobic deep pocket of *Blys* would lead us to search for hydrophobic side chains at position R5, that pocket (Figure 5-6 and Figure 5-8) has a strong hydrogen bond acceptor (the main chain oxygen of Cys 232). The LIDAEUS experiments have shown that that this oxygen can be exploited to strengthen the interaction of Blys with a bound ligand in that pocket. Compared to the size of the pocket, Leu is relatively small and we would like to test larger hydrophobic side chains in that pocket as well (and possibly with variations that would include a hydrogen bond donor). R4 was to be

given similar hydrophobic choices, while R3 and R6 would be given polar choices, mimicking the Asp and Arg interactions with *Blys* (Figure 5-16).

Table 5-4: Side chain Choices for residues R3, R4, R5, and R6 in the first docking experiment (see Figure 5-19 and Figure 5-20).

R3	R4	R5	R6
Asp	Tbu	Tbu	D-Arg
Glu	Cha	Cha	D-Hrg
Hgu	Gln	Fgl	D-Lys
Pcf	Leu	Phe	
Mcf	Dmb	Tza	
Ocf	Pza	Pza	
Ttz	Cpa	Dmb	
	Htr	His	
	Tfa	Val	

The choices of the first experiment are given in Table 5-4, and a 2 dimensional description of the non-natural side chains is found in Figure 5-19 and Figure 5-20. With R3 given 7 choices, R4 and R5 each given 9 choices, and R6 given 3 choices, that sums up to a total of $7 \times 9 \times 9 \times 3 = 1701$ molecules. These molecules were docked with Autodock using standard parameters (except for the GA_RUN being set to 3). GA_RUN sets the number of starting points for the Monte-Carlo genetic algorithm. The more starting points you have, the higher the probability to find the global minimum in your energy minimization attempts; the corollary to this is the longer your experiments will take. As this docking run is large (1701 peptide molecules with a large number of rotatable bonds), the GA_RUN was fixed to 3 instead of 10.

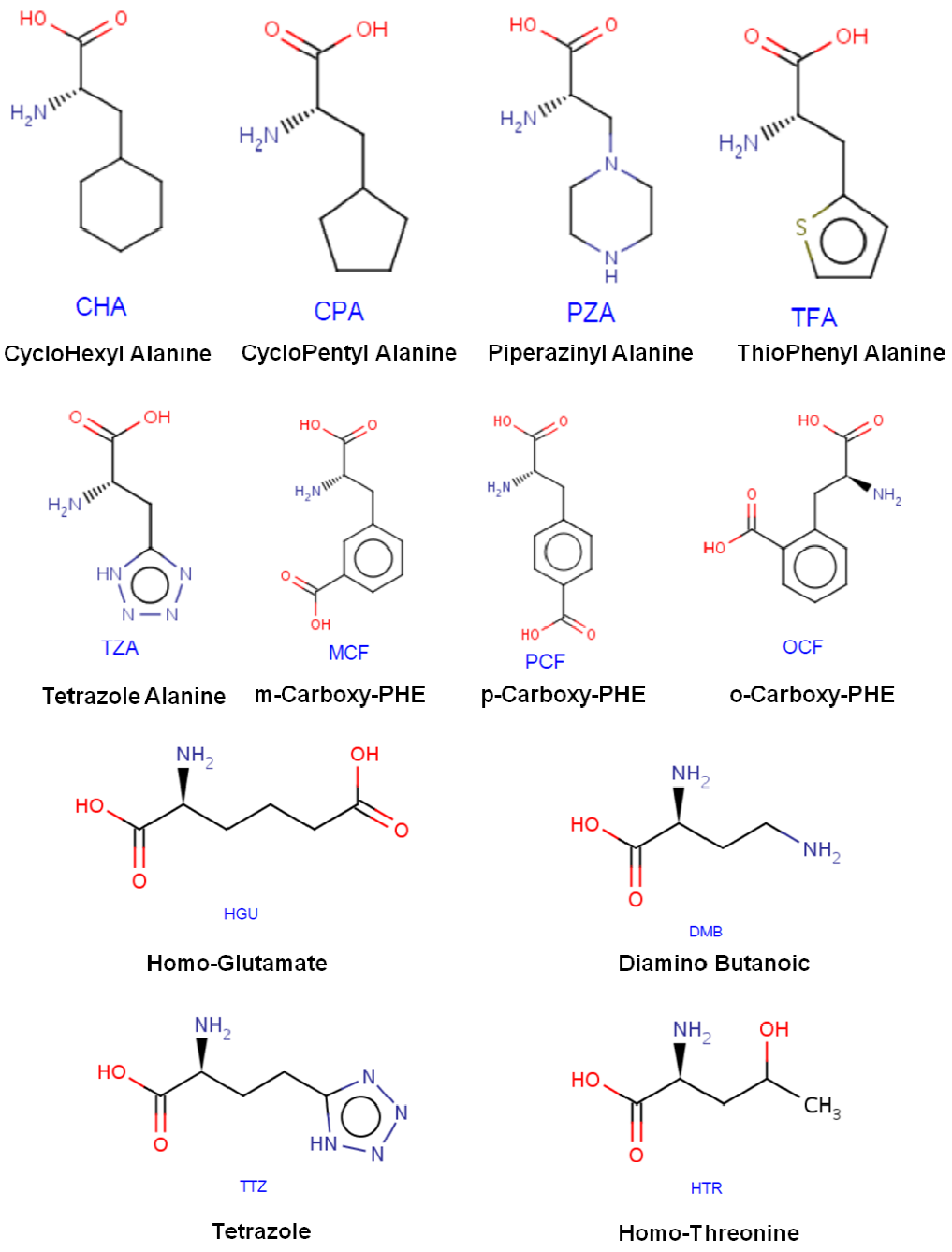

Figure 5-19: Two-dimensional representation of the nonnatural side chains used (Part 1/2).

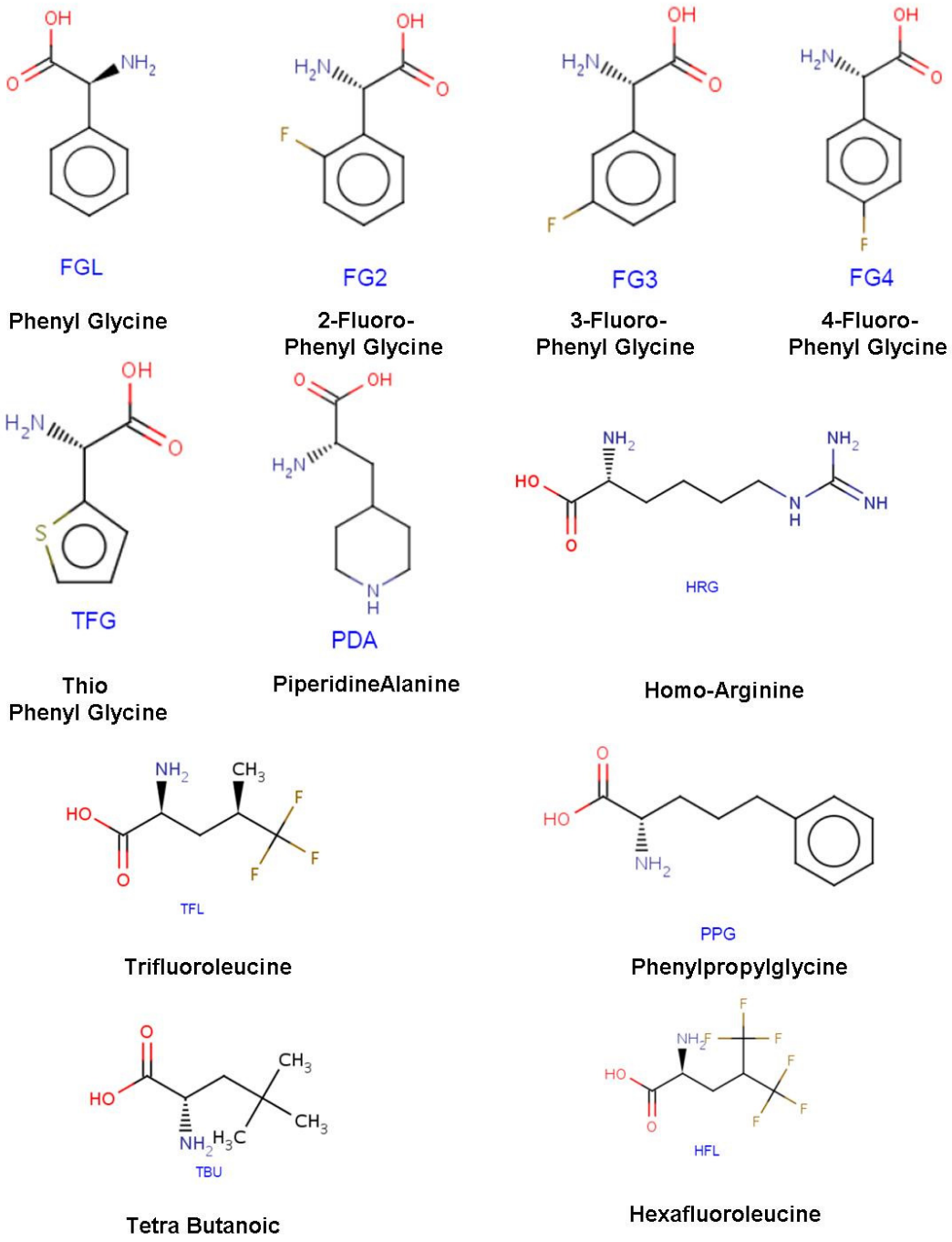

Figure 5-20: Two-dimensional representation of the nonnatural side chains used (Part 2/2).

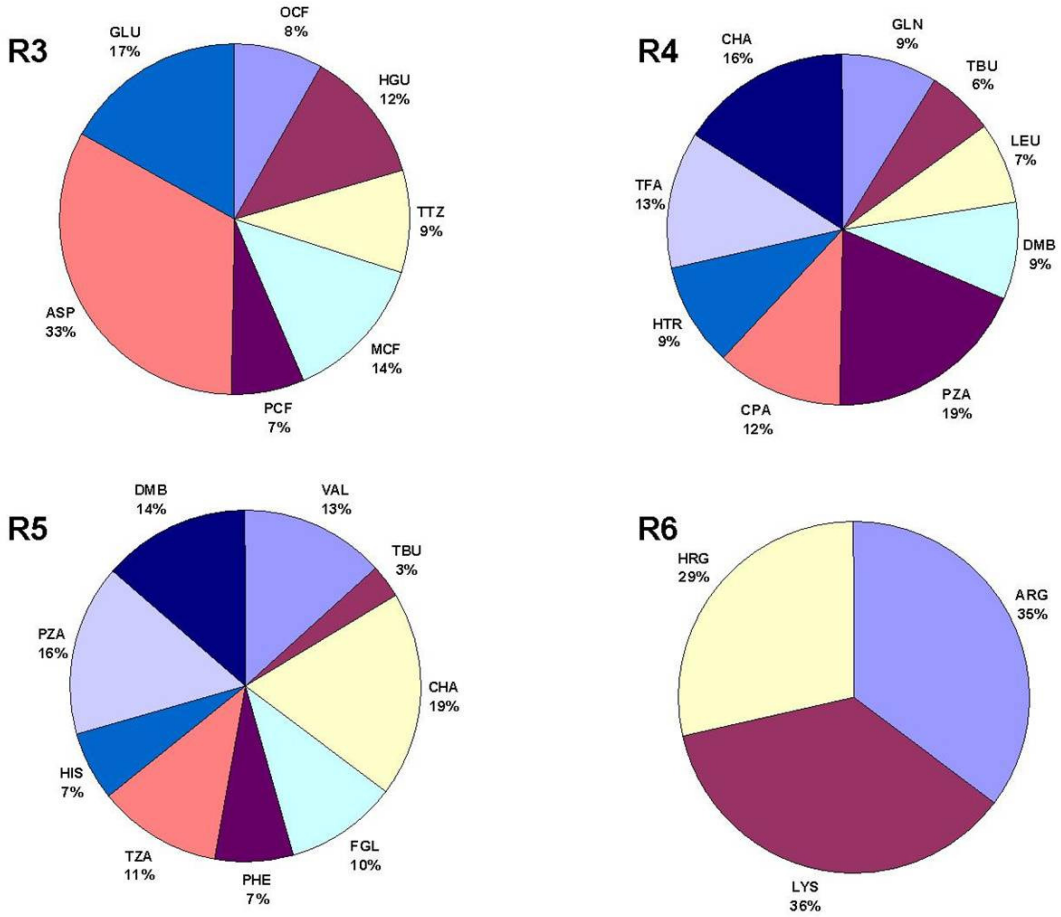

Figure 5-21: The distribution of the side chain choices in the correct docking poses from the first docking experiment.

Out of the 1701 molecules tested, 307 (18%) exhibited the correct docking pose (where the backbone of the docked hexapeptide is mimicking the position of the backbone of the original *BaffR* ligand. Autodock generated energy scores for these 1701 poses where in the range of [-8.79, -2.12] kcal/mol. The best poses were studied to evaluate the side chain compositons at R3-R6, and the results are shown in Figure 5-21. R3 showed a strong preference for Asp. R4 showed less preference for Dmb, Leu, Tbu, Gln, and Htr (when compared to the other choices). In R5, Phe was the only side chain to be considered as disfavored by the docking algorithm. No

preference was shown in R6, but Lys was excluded as since it did not contribute much to the interaction energy of the poses.

Table 5-5: Side chain Choices for residues R4, R5, and R6 in the second experiment.

R3	R4	R5	R6
Asp	His	Val	D-Arg
	Arg	Leu	D-Hrg
	Leu	Cha	
	Lys	Fgl	
	Pza	Fg2	
	Ppg	Fg3	
	Trp	Fg4	
	Tfa	Trp	
	Cha	Tza	
		His	
		Pza	
		Pda	
		Cpa	
		Tfg	
		Hfl	
		Tfl	
		Tbu	

Consequently, a second docking run was designed (this time with GA_RUN set to 10). R3 was fixed to Asp. R6 was fixed to Arg and Hrg. R4 and R5 were given 9 and 17 choices respectively (the original choices without the disfavored side chains, and a few extra choices that include Fluorine atoms in the side chains (Table 5-5). A total of $1 \times 9 \times 17 \times 2 = 306$ molecules were tested. 217 (71%) of them were docked in the correct pose. the Autodock energy scores were in the range of [-8.07, -3.03] kcal/mol. This indicates an increase in the quality of the choices as a larger percentage is being docked in the correct pose, and the high end of the energy score interval has become more negative. Although the low end of the energy scores has shifted a bit to the right (which is not optimal), it is still considered as a good result

as this range represents a majority of correct poses (unlike the first experiment where the majority of the poses were not mimicking the original conformation). Figure 5-22 shows the distribution of the top 102 of the correct poses. The library being tested is made up of permutations of side chains on R3-R6. With 71% of the docked molecules were in a correct pose, searching for the favored side chain choices at R3-R6 among these poses will not generate a strong signal, and the choices will tend to be equally distributed. Therefore the top third of the docking results were chosen (102 structures) to give a better signal of favored side chains. Based on these results, Leu, Arg, and Lys were removed as possibilities for R4 while Leu, Pza, Hfl, Tbu, Val, Tza, Tfl, His, and Trp were removed as possibilities for R5. Although Hrg was not as favorable as Arg in R6, the next experiment was to increase choices in R6 and Hrg was left as a reference to compare the new results with.

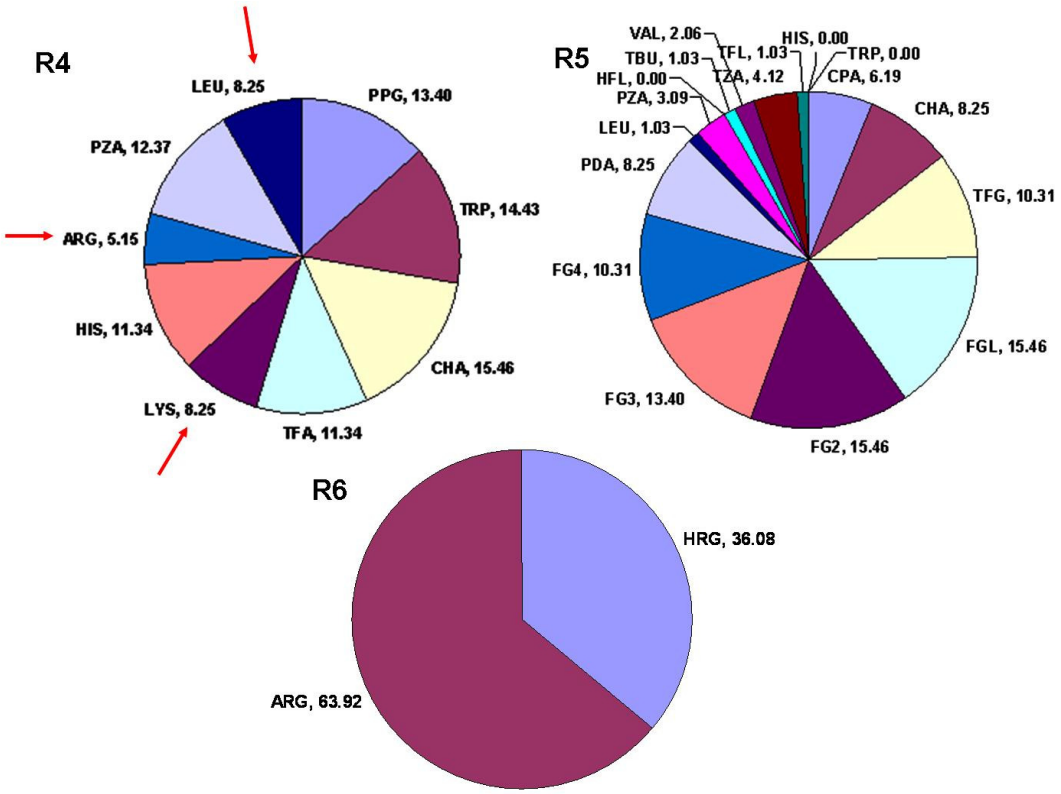

Figure 5-22: The distribution of the side chain choices in the top 102 correct docking poses from the second docking experiment.

Table 5-6: Side chain Choices for residues R4, R5, and R6 in the third experiment.

R3	R4	R5	R6
Asp	His	Cha	D-Arg
	Pza	Fgl	D-Hrg
	Ppg	Fg2	D-Tfa
	Trp	Fg3	D-Phe
	Tfa	Fg4	D-Leu
	Cha	Pda	L-Leu
		Cpa	L-Phe
		Tfg	L-Tfa
			L-Cpa
			L-His
			L-Asn
			L-Tbu
			L-Val

The third docking run was designed to test for further possibilities on R6 and included introducing L-Aminoacids to R6. As per Table 5-6, a total of 1 × 6 × 8 × 13 = 624 molecules were designed and tested. 534 (85.6%) of these molecules were in the correct pose. the Autodock energy scores were in the range of [-8.54, -4.53] kcal/mol. With the percentage of docked molecules exhibiting the same pose as the original ligand and the range of the energy scores shifting to the left on both of its extremities, the docking run was considered successful. Figure 5-23 shows the distribution of the side chain choices in the top 208 of the correct poses (again, taking 33%). As a result, Tfa and His were discarded from R4 and Fg4 was discarded from R5. As for R6, D-Tfa, D-Phe, L-Phe, L-Tfa, L-Cpa, L-His, and L-Tbu were also discarded.

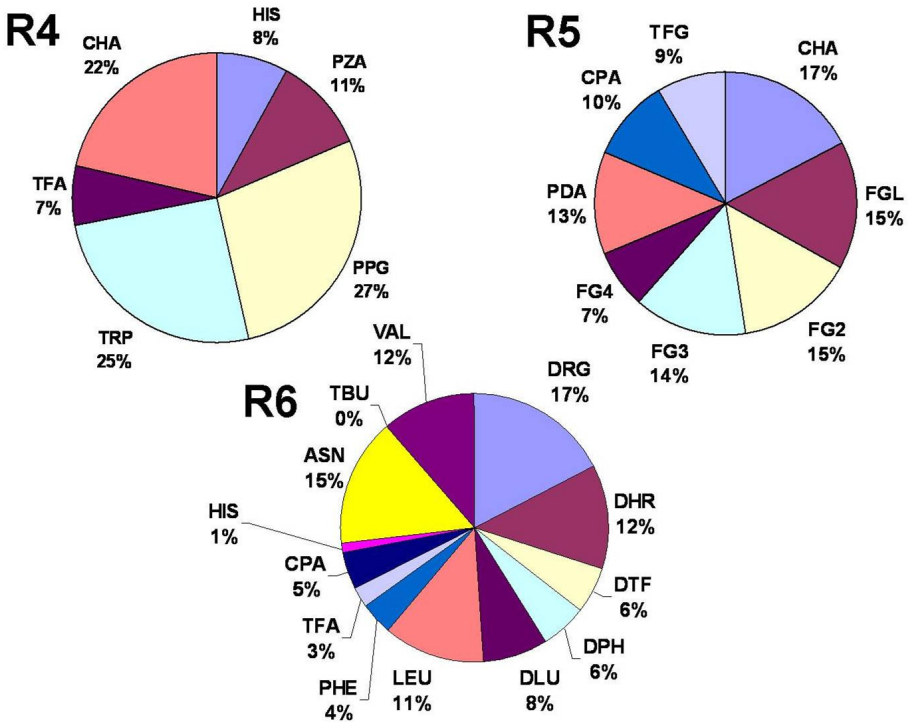

Figure 5-23: The distribution of the side chain choices in the top 208 correct docking poses from the third docking experiment.

Table 5-7: Side chain Choices for residues R3, R4, R5, and R6 in the fourth experiment.

R3	R4	R5	R6
Asp	Pza	Cha	D-Arg
Glu	Ppg	Fgl	D-Hrg
Hgu	Trp	Fg2	D-Leu
	Cha	Fg3	L-Leu
		Pda	L-Asn
		Cpa	L-Val
		Tfg	

A final docking run to check for further optimization of R3 was designed. Glu and Hgu were taken (based on the results of the first docking experiment) and tested again at the fine resolution of GA_RUN = 10. It was already clear that Asp was the best choice for R3, but we wanted to investigate whether Glu and Hgu (which would actually produce the same interactions like Asp) would also improve the docking energies. As per Table 5-7, a total of $3 \times 4 \times 7 \times 6 = 504$ molecules were designed and tested. 460 (91%) of them were docked in the correct pose, and the Autodock energy scores were in the range of [-8.90, -4.64] kcal/mol; showing further improvement on the side chain choices. Figure 5-24 shows the distribution of the top 168 of the correct poses (33% of the 504 choices). Although Glu and Hgu are not as favorable as Asp, they did enhance the Autodock energy scores and are worth trying *in vitro*. R4 and R5 show no favoring at all between the choices, asserting that no further improvements could be made with the current choices. Same observation goes for R6, except for D-Leu being less favored than everything else. The details of the entire set of docked molecules from the fourth docking experiment are found in Table 5-8.

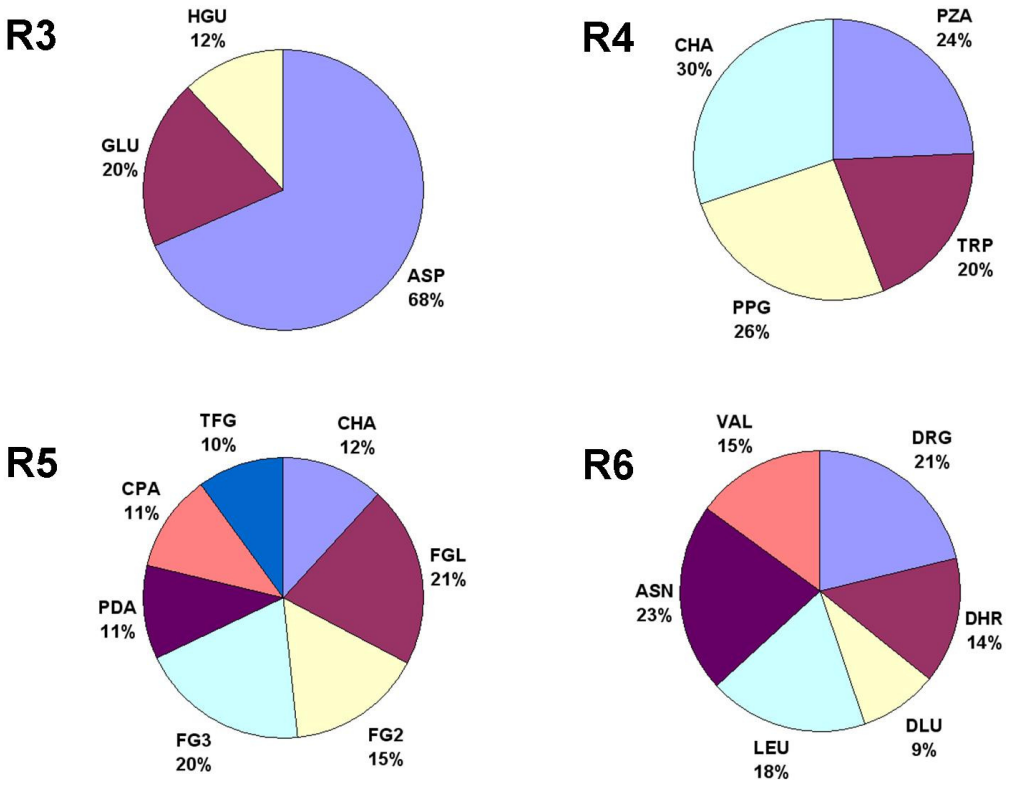

Figure 5-24: The distribution of the side chain choices in the top 168 correct docking poses from the fourth docking experiment.

With those four Autodock experiments, we were capable of converging to the best possible choices of aminoacids that could be used to synthesize molecules that would inhibit the *Blys*/*BaffR* complex. At every step, we discarded the unfavorable choices and tried to formulate educated guesses at what new side chains would help the interaction become stronger (as long as the side chains are synthesizable as well).

Table 5-8: The cyclic hexapeptide virtual screening library. The number of clusters denotes the number of clusters in which the 10 docking experiments per molecule can be classified (the less the clusters, the better the confidence). Drg, Dhr, and Dlu refer to D-Arg, D-Hrg, and D-Leu respectively.

R3	R4	R5	R6	Predicted Energy (kcal/mol)	Predicted Affinity	Number of Clusters
Asp	Ppg	Cha	Drg	-8.9	301.34nM	5
Asp	Ppg	Fgl	Drg	-8.51	580.94nM	4
Asp	Trp	Fgl	Drg	-8.49	595.70nM	5
Asp	Trp	Fg3	Drg	-8.48	608.00nM	3
Asp	Cha	Pda	Leu	-8.37	737.37nM	2
Asp	Ppg	Fg3	Drg	-8.34	769.30nM	1
Asp	Cha	Fg3	Drg	-8.32	793.35nM	3
Asp	Ppg	Pda	Asn	-8.32	790.98nM	5
Asp	Ppg	Fgl	Asn	-8.26	887.65nM	1
Asp	Cha	Fg2	Drg	-8.24	904.47nM	3
Asp	Trp	Pda	Val	-8.21	958.52nM	2
Asp	Ppg	Cha	Val	-8.21	965.66nM	4
Asp	Ppg	Fg3	Asn	-8.21	957.56nM	3
Asp	Cha	Pda	Val	-8.18	1.01µM	2
Asp	Trp	Fgl	Dhr	-8.14	1.09µM	4
Asp	Trp	Pda	Dhr	-8.09	1.17µM	8
Asp	Ppg	Fgl	Leu	-8.07	1.22µM	3
Asp	Cha	Cha	Dhr	-8.01	1.35µM	7
Asp	Trp	Fg3	Asn	-7.98	1.42µM	2
Asp	Ppg	Fg2	Asn	-7.97	1.43µM	2
Asp	Ppg	Cha	Leu	-7.95	1.49µM	2
Asp	Ppg	Fg2	Val	-7.95	1.49µM	2
Asp	Trp	Fg3	Leu	-7.94	1.51µM	3
Asp	Ppg	Fg2	Drg	-7.94	1.52µM	3
Asp	Ppg	Fg3	Dhr	-7.94	1.52µM	5
Asp	Ppg	Tfg	Dhr	-7.94	1.53µM	4
Asp	Pza	Cha	Dhr	-7.94	1.52µM	4
Asp	Ppg	Cha	Dlu	-7.93	1.53µM	5
Asp	Ppg	Cpa	Asn	-7.93	1.53µM	4
Asp	Cha	Fgl	Drg	-7.92	1.56µM	2
Asp	Trp	Fg2	Drg	-7.91	1.60µM	4
Glu	Trp	Fg3	Dhr	-7.91	1.59µM	5
Asp	Ppg	Pda	Val	-7.91	1.60µM	5
Asp	Trp	Pda	Leu	-7.9	1.62µM	7
Asp	Ppg	Fgl	Dlu	-7.89	1.63µM	1
Asp	Ppg	Pda	Leu	-7.89	1.66µM	5
Asp	Trp	Fgl	Asn	-7.88	1.68µM	2
Asp	Trp	Pda	Asn	-7.88	1.68µM	3
Asp	Ppg	Fg3	Leu	-7.88	1.68µM	2
Asp	Trp	Cha	Asn	-7.86	1.73µM	5
Asp	Cha	Tfg	Dhr	-7.86	1.73µM	5

Asp	Cha	Fg2	Dhr	-7.84	1.79μM	4
Asp	Pza	Fgl	Dhr	-7.83	1.82μM	2
Glu	Cha	Fg2	Dhr	-7.83	1.83μM	4
Asp	Cha	Pda	Asn	-7.82	1.86μM	3
Asp	Ppg	Cha	Asn	-7.82	1.87μM	4
Asp	Trp	Cha	Val	-7.8	1.92μM	4
Asp	Pza	Fg2	Drg	-7.79	1.96μM	3
Glu	Ppg	Fgl	Asn	-7.79	1.95μM	3
Asp	Trp	Fg2	Asn	-7.78	1.98μM	3
Asp	Cha	Fg3	Asn	-7.77	2.02μM	1
Asp	Ppg	Fgl	Dhr	-7.75	2.09μM	3
Asp	Cha	Fg3	Leu	-7.74	2.13μM	3
Asp	Trp	Cpa	Asn	-7.73	2.17μM	2
Asp	Cha	Fgl	Asn	-7.73	2.17μM	1
Asp	Pza	Fgl	Drg	-7.71	2.22μM	1
Asp	Cha	Cpa	Val	-7.7	2.26μM	3
Asp	Pza	Tfg	Drg	-7.68	2.34μM	4
Glu	Ppg	Fg3	Drg	-7.66	2.41μM	4
Asp	Cha	Fgl	Leu	-7.66	2.43μM	2
Asp	Ppg	Fg2	Leu	-7.66	2.42μM	4
Asp	Cha	Cpa	Leu	-7.65	2.48μM	3
Hgu	Cha	Cpa	Asn	-7.65	2.45μM	4
Asp	Cha	Fg2	Asn	-7.64	2.53μM	2
Asp	Trp	Cpa	Val	-7.64	2.52μM	3
Asp	Cha	Fg3	Dhr	-7.63	2.55μM	1
Hgu	Cha	Fg3	Asn	-7.63	2.54μM	5
Asp	Trp	Cha	Dlu	-7.62	2.58μM	5
Glu	Pza	Fgl	Dhr	-7.62	2.59μM	6
Hgu	Pza	Pda	Leu	-7.62	2.60μM	6
Asp	Ppg	Cpa	Val	-7.61	2.64μM	3
Asp	Cha	Fgl	Dhr	-7.6	2.66μM	4
Hgu	Pza	Fg2	Dhr	-7.6	2.69μM	6
Asp	Trp	Cpa	Drg	-7.59	2.75μM	5
Asp	Trp	Cpa	Leu	-7.59	2.71μM	4
Asp	Trp	Fg2	Leu	-7.59	2.74μM	5
Glu	Ppg	Fg3	Asn	-7.59	2.71μM	5
Asp	Ppg	Cpa	Drg	-7.58	2.78μM	6
Asp	Pza	Pda	Val	-7.57	2.81μM	2
Asp	Trp	Fgl	Leu	-7.57	2.83μM	5
Asp	Ppg	Fg3	Val	-7.57	2.84μM	4
Asp	Trp	Pda	Drg	-7.55	2.92μM	7
Glu	Ppg	Tfg	Dhr	-7.55	2.90μM	8
Glu	Pza	Tfg	Drg	-7.55	2.94μM	4
Hgu	Cha	Fg3	Dhr	-7.55	2.92μM	6
Hgu	Trp	Fg3	Drg	-7.55	2.94μM	2
Asp	Cha	Cpa	Asn	-7.55	2.91μM	2
Asp	Pza	Cha	Val	-7.55	2.92μM	2
Asp	Trp	Fg3	Dhr	-7.52	3.09μM	4

Hgu	Pza	Fgl	Val	-7.51	3.15μM	2
Asp	Pza	Cha	Asn	-7.51	3.15μM	2
Asp	Trp	Fg3	Val	-7.5	3.18μM	4
Glu	Pza	Fg2	Dhr	-7.5	3.21μM	7
Hgu	Cha	Tfg	Drg	-7.5	3.16μM	3
Asp	Pza	Cha	Dlu	-7.5	3.17μM	2
Asp	Pza	Pda	Leu	-7.49	3.22μM	2
Hgu	Pza	Fgl	Asn	-7.49	3.21μM	4
Asp	Trp	Tfg	Leu	-7.47	3.34μM	5
Glu	Trp	Fg3	Drg	-7.47	3.37μM	5
Hgu	Cha	Pda	Leu	-7.47	3.34μM	6
Asp	Pza	Fgl	Leu	-7.46	3.43μM	2
Glu	Pza	Cha	Dhr	-7.46	3.40μM	5
Glu	Cha	Cpa	Leu	-7.45	3.45μM	6
Glu	Cha	Fgl	Leu	-7.45	3.46μM	4
Hgu	Cha	Fg2	Val	-7.45	3.45μM	4
Asp	Trp	Tfg	Asn	-7.44	3.51μM	3
Asp	Ppg	Fg2	Dlu	-7.43	3.61μM	4
Asp	Trp	Cha	Leu	-7.42	3.66μM	8
Asp	Cha	Pda	Drg	-7.42	3.61μM	6
Asp	Cha	Tfg	Asn	-7.42	3.62μM	1
Asp	Ppg	Fg3	Dlu	-7.42	3.63μM	4
Asp	Trp	Tfg	Drg	-7.4	3.74μM	3
Asp	Cha	Cha	Val	-7.4	3.77μM	1
Glu	Cha	Cpa	Drg	-7.39	3.84μM	5
Hgu	Cha	Fgl	Val	-7.39	3.84μM	4
Asp	Pza	Cpa	Leu	-7.39	3.81μM	3
Asp	Pza	Fg3	Asn	-7.38	3.90μM	1
Asp	Pza	Fg2	Asn	-7.38	3.89μM	1
Asp	Pza	Cpa	Asn	-7.37	3.95μM	1
Asp	Cha	Cpa	Drg	-7.36	4.00μM	4
Glu	Ppg	Cha	Val	-7.36	4.00μM	7
Asp	Cha	Tfg	Drg	-7.36	4.01μM	3
Asp	Ppg	Tfg	Asn	-7.36	4.00μM	2
Glu	Pza	Cpa	Asn	-7.35	4.07μM	2
Hgu	Ppg	Fgl	Drg	-7.34	4.16μM	8
Asp	Pza	Fg3	Val	-7.33	4.25μM	2
Asp	Pza	Fgl	Asn	-7.33	4.20μM	1
Hgu	Cha	Cpa	Drg	-7.33	4.22μM	7
Glu	Cha	Fgl	Val	-7.32	4.30μM	3
Asp	Cha	Fg2	Leu	-7.31	4.42μM	5
Asp	Cha	Fg2	Val	-7.31	4.41μM	2
Glu	Ppg	Fgl	Dhr	-7.31	4.41μM	3
Glu	Ppg	Pda	Drg	-7.31	4.36μM	6
Asp	Pza	Pda	Asn	-7.3	4.47μM	2
Asp	Pza	Cpa	Val	-7.3	4.49μM	2
Asp	Cha	Fg2	Dlu	-7.29	4.56μM	2
Glu	Cha	Fg2	Drg	-7.29	4.54μM	4

Glu	Cha	Fgl	Asn	-7.29	4.56μM	2
Asp	Trp	Fg2	Val	-7.28	4.63μM	5
Asp	Cha	Tfg	Leu	-7.28	4.60μM	2
Asp	Pza	Tfg	Dhr	-7.27	4.72μM	4
Asp	Cha	Fgl	Dlu	-7.27	4.68μM	4
Glu	Trp	Fg3	Asn	-7.27	4.67μM	5
Asp	Cha	Cha	Dlu	-7.27	4.73μM	3
Hgu	Pza	Fg3	Asn	-7.27	4.67μM	3
Asp	Ppg	Fgl	Val	-7.26	4.73μM	3
Glu	Ppg	Fg2	Asn	-7.25	4.88μM	5
Hgu	Pza	Fgl	Leu	-7.25	4.82μM	4
Asp	Pza	Fg2	Dhr	-7.25	4.84μM	5
Glu	Pza	Fg3	Leu	-7.23	4.99μM	5
Asp	Trp	Pda	Dlu	-7.22	5.08μM	2
Glu	Pza	Fgl	Leu	-7.22	5.07μM	4
Hgu	Cha	Pda	Val	-7.21	5.17μM	5
Hgu	Ppg	Fg3	Drg	-7.21	5.21μM	6
Asp	Ppg	Tfg	Drg	-7.21	5.17μM	1
Glu	Cha	Fgl	Drg	-7.2	5.31μM	2
Asp	Cha	Fg3	Dlu	-7.2	5.26μM	1
Glu	Pza	Fg2	Leu	-7.2	5.24μM	3
Hgu	Pza	Cpa	Leu	-7.2	5.28μM	6
Glu	Ppg	Fgl	Val	-7.19	5.35μM	4
Hgu	Cha	Fg3	Drg	-7.19	5.34μM	3
Glu	Pza	Cha	Val	-7.18	5.43μM	1
Glu	Pza	Fg3	Drg	-7.18	5.47μM	4
Glu	Pza	Fgl	Val	-7.18	5.46μM	5
Asp	Pza	Tfg	Asn	-7.17	5.54μM	1
Asp	Pza	Fg2	Leu	-7.16	5.62μM	4
Asp	Trp	Fg3	Dlu	-7.16	5.67μM	4
Asp	Trp	Fgl	Dlu	-7.16	5.68μM	4
Glu	Cha	Tfg	Dhr	-7.16	5.65μM	6
Asp	Cha	Pda	Dlu	-7.16	5.60μM	3
Asp	Cha	Cha	Leu	-7.16	5.69μM	4
Glu	Cha	Fg3	Drg	-7.15	5.74μM	5
Asp	Ppg	Pda	Dlu	-7.15	5.70μM	5
Asp	Pza	Fg3	Dlu	-7.14	5.83μM	4
Glu	Cha	Fg3	Asn	-7.14	5.79μM	2
Glu	Cha	Fgl	Dlu	-7.14	5.85μM	4
Asp	Pza	Fg2	Val	-7.13	5.95μM	1
Asp	Trp	Tfg	Dhr	-7.13	5.90μM	4
Hgu	Cha	Cha	Val	-7.13	5.93μM	6
Glu	Ppg	Fgl	Leu	-7.12	6.02μM	4
Asp	Ppg	Tfg	Dlu	-7.12	6.07μM	2
Asp	Pza	Fg3	Drg	-7.11	6.15μM	1
Glu	Ppg	Fgl	Drg	-7.11	6.10μM	3
Glu	Ppg	Tfg	Drg	-7.11	6.12μM	5
Hgu	Cha	Fgl	Dlu	-7.11	6.15μM	5

Hgu	Pza	Tfg	Drg	-7.11	6.10μM	3
Asp	Trp	Cpa	Dhr	-7.1	6.22μM	5
Glu	Cha	Pda	Leu	-7.1	6.26μM	3
Asp	Cha	Cha	Asn	-7.09	6.33μM	2
Asp	Cha	Fgl	Val	-7.09	6.33μM	1
Hgu	Trp	Fg3	Dhr	-7.09	6.40μM	7
Asp	Pza	Cha	Leu	-7.09	6.36μM	4
Hgu	Cha	Fg3	Val	-7.08	6.41μM	4
Hgu	Trp	Fg3	Val	-7.08	6.44μM	6
Glu	Cha	Fgl	Dhr	-7.07	6.53μM	6
Glu	Ppg	Cha	Asn	-7.07	6.57μM	7
Glu	Ppg	Fg2	Drg	-7.07	6.52μM	8
Glu	Trp	Fg2	Drg	-7.07	6.57μM	5
Hgu	Cha	Pda	Drg	-7.07	6.59μM	7
Glu	Cha	Pda	Asn	-7.06	6.67μM	4
Glu	Ppg	Cpa	Val	-7.06	6.65μM	2
Hgu	Ppg	Tfg	Dhr	-7.06	6.68μM	9
Hgu	Trp	Fg2	Drg	-7.06	6.69μM	4
Hgu	Trp	Fgl	Drg	-7.06	6.67μM	7
Glu	Cha	Pda	Val	-7.05	6.78μM	6
Glu	Ppg	Fg2	Dlu	-7.05	6.77μM	4
Asp	Pza	Pda	Drg	-7.04	6.88μM	4
Asp	Cha	Fg3	Val	-7.04	6.88μM	2
Hgu	Cha	Fg2	Dlu	-7.04	6.92μM	3
Hgu	Ppg	Fgl	Dlu	-7.04	6.89μM	3
Asp	Ppg	Tfg	Val	-7.04	6.87μM	2
Asp	Cha	Tfg	Dlu	-7.03	6.97μM	2
Glu	Ppg	Pda	Dhr	-7.02	7.20μM	10
Glu	Pza	Fg3	Asn	-7.02	7.15μM	4
Asp	Trp	Fg2	Dhr	-7.01	7.24μM	6
Glu	Ppg	Cpa	Dhr	-7.01	7.23μM	7
Glu	Pza	Fg2	Val	-7.01	7.32μM	3
Glu	Trp	Cha	Dlu	-7.01	7.23μM	5
Asp	Trp	Cha	Drg	-7	7.42μM	4
Asp	Trp	Fgl	Val	-7	7.34μM	4
Hgu	Pza	Cpa	Asn	-7	7.44μM	6
Hgu	Pza	Pda	Val	-7	7.36μM	6
Asp	Ppg	Pda	Dhr	-7	7.38μM	8
Asp	Trp	Fg2	Dlu	-6.99	7.51μM	3
Glu	Trp	Fgl	Leu	-6.99	7.58μM	4
Hgu	Trp	Cpa	Asn	-6.99	7.51μM	8
Glu	Cha	Fg3	Val	-6.97	7.83μM	3
Glu	Trp	Cpa	Asn	-6.97	7.81μM	4
Hgu	Trp	Fg3	Asn	-6.97	7.75μM	3
Hgu	Trp	Tfg	Val	-6.97	7.75μM	3
Asp	Pza	Fgl	Val	-6.96	7.94μM	1
Glu	Pza	Fg3	Val	-6.96	7.93μM	5
Glu	Trp	Fg2	Dlu	-6.96	7.88μM	3

Hgu	Ppg	Cha	Drg	-6.96	7.87μM	6
Asp	Ppg	Cpa	Dlu	-6.96	7.90μM	6
Asp	Pza	Fg3	Leu	-6.95	8.05μM	4
Asp	Cha	Pda	Dhr	-6.95	8.10μM	4
Asp	Cha	Cpa	Dlu	-6.95	8.04μM	2
Glu	Trp	Fg2	Dhr	-6.94	8.12μM	5
Asp	Pza	Cpa	Dhr	-6.93	8.34μM	4
Glu	Pza	Pda	Val	-6.92	8.40μM	5
Glu	Trp	Cpa	Leu	-6.92	8.43μM	6
Hgu	Ppg	Fgl	Asn	-6.92	8.45μM	4
Asp	Pza	Fgl	Dlu	-6.91	8.65μM	1
Asp	Trp	Tfg	Val	-6.91	8.64μM	5
Glu	Trp	Cpa	Val	-6.91	8.66μM	6
Asp	Cha	Cha	Drg	-6.91	8.66μM	3
Glu	Pza	Pda	Leu	-6.9	8.70μM	4
Glu	Trp	Pda	Leu	-6.9	8.75μM	5
Hgu	Cha	Fgl	Drg	-6.9	8.69μM	3
Hgu	Pza	Tfg	Leu	-6.9	8.71μM	7
Hgu	Pza	Fgl	Dhr	-6.89	8.90μM	6
Asp	Pza	Fg3	Dhr	-6.87	9.24μM	6
Glu	Cha	Fg2	Asn	-6.87	9.20μM	5
Glu	Ppg	Fg2	Dhr	-6.87	9.28μM	6
Hgu	Trp	Pda	Leu	-6.87	9.19μM	7
Glu	Trp	Pda	Dhr	-6.86	9.35μM	7
Glu	Pza	Fg2	Drg	-6.85	9.56μM	3
Glu	Trp	Fg2	Asn	-6.85	9.59μM	4
Glu	Trp	Tfg	Asn	-6.85	9.55μM	5
Asp	Pza	Cpa	Drg	-6.85	9.50μM	3
Glu	Pza	Cpa	Leu	-6.84	9.71μM	3
Glu	Pza	Fg3	Dlu	-6.84	9.74μM	2
Hgu	Cha	Fg2	Asn	-6.84	9.66μM	3
Hgu	Ppg	Fg3	Val	-6.84	9.69μM	6
Glu	Ppg	Cha	Dlu	-6.83	9.86μM	8
Glu	Ppg	Tfg	Asn	-6.83	9.79μM	4
Hgu	Pza	Cha	Val	-6.83	9.87μM	5
Asp	Pza	Pda	Dlu	-6.82	10.03μM	4
Glu	Cha	Cpa	Val	-6.82	10.10μM	2
Glu	Ppg	Fg2	Val	-6.82	10.08μM	3
Glu	Ppg	Fgl	Dlu	-6.82	10.08μM	4
Asp	Ppg	Cha	Dhr	-6.82	10.03μM	6
Hgu	Trp	Fgl	Val	-6.82	10.03μM	4
Glu	Pza	Cha	Leu	-6.81	10.17μM	5
Hgu	Cha	Cha	Leu	-6.81	10.22μM	5
Hgu	Trp	Fgl	Dhr	-6.81	10.11μM	5
Hgu	Cha	Cpa	Leu	-6.8	10.37μM	4
Asp	Cha	Tfg	Val	-6.8	10.45μM	3
Hgu	Pza	Fg2	Dlu	-6.8	10.32μM	3
Hgu	Trp	Fg2	Val	-6.8	10.28μM	5

Asp	Trp	Cpa	Dlu	-6.79	10.51μM	4
Glu	Cha	Fg2	Dlu	-6.79	10.55μM	4
Glu	Ppg	Fg3	Val	-6.79	10.48μM	4
Glu	Ppg	Pda	Val	-6.78	10.75μM	4
Glu	Ppg	Tfg	Val	-6.78	10.68μM	3
Glu	Pza	Tfg	Asn	-6.78	10.67μM	2
Glu	Trp	Fg3	Val	-6.78	10.65μM	3
Hgu	Pza	Cpa	Dlu	-6.78	10.76μM	7
Hgu	Pza	Fg2	Drg	-6.78	10.80μM	5
Asp	Trp	Tfg	Dlu	-6.77	10.96μM	2
Glu	Cha	Tfg	Leu	-6.77	10.85μM	6
Glu	Pza	Cpa	Dlu	-6.77	10.91μM	3
Asp	Pza	Tfg	Leu	-6.76	11.04μM	2
Hgu	Ppg	Fg2	Drg	-6.76	11.08μM	7
Asp	Pza	Tfg	Val	-6.75	11.33μM	2
Glu	Ppg	Fg3	Dlu	-6.75	11.29μM	7
Glu	Ppg	Pda	Dlu	-6.75	11.25μM	8
Glu	Pza	Fg2	Asn	-6.75	11.27μM	2
Glu	Trp	Fgl	Dlu	-6.75	11.24μM	3
Hgu	Ppg	Fg3	Asn	-6.75	11.24μM	4
Hgu	Ppg	Pda	Asn	-6.75	11.36μM	9
Hgu	Pza	Fg3	Drg	-6.75	11.25μM	5
Glu	Cha	Fg2	Val	-6.74	11.39μM	5
Glu	Cha	Tfg	Drg	-6.74	11.55μM	4
Glu	Trp	Pda	Val	-6.74	11.56μM	5
Hgu	Cha	Tfg	Dlu	-6.74	11.44μM	5
Glu	Cha	Cpa	Asn	-6.73	11.58μM	3
Glu	Trp	Cha	Val	-6.73	11.65μM	7
Glu	Trp	Fg2	Val	-6.73	11.66μM	4
Glu	Ppg	Pda	Asn	-6.72	11.79μM	6
Asp	Ppg	Fg2	Dhr	-6.72	11.91μM	5
Glu	Cha	Fg3	Dlu	-6.71	12.07μM	5
Glu	Trp	Pda	Drg	-6.71	12.13μM	7
Glu	Cha	Tfg	Asn	-6.7	12.25μM	4
Glu	Pza	Fgl	Asn	-6.7	12.33μM	3
Hgu	Pza	Cha	Asn	-6.7	12.32μM	3
Hgu	Trp	Pda	Asn	-6.7	12.36μM	7
Glu	Cha	Cpa	Dlu	-6.69	12.45μM	4
Glu	Cha	Pda	Drg	-6.69	12.43μM	6
Hgu	Ppg	Fg3	Dlu	-6.69	12.49μM	8
Hgu	Pza	Cha	Drg	-6.69	12.53μM	7
Hgu	Trp	Fg2	Dlu	-6.69	12.47μM	7
Hgu	Ppg	Fgl	Val	-6.68	12.74μM	5
Hgu	Pza	Fg3	Dlu	-6.68	12.66μM	6
Glu	Cha	Fg2	Leu	-6.67	12.96μM	4
Glu	Cha	Pda	Dhr	-6.67	12.86μM	6
Glu	Ppg	Cha	Leu	-6.67	12.81μM	2
Glu	Trp	Fgl	Asn	-6.67	12.89μM	4

Glu	Trp	Fgl	Drg	-6.67	12.96μM	4
Hgu	Cha	Fgl	Asn	-6.67	12.89μM	6
Hgu	Trp	Fgl	Dlu	-6.67	12.83μM	6
Glu	Ppg	Fg3	Leu	-6.66	13.20μM	5
Glu	Trp	Fgl	Val	-6.66	13.08μM	4
Hgu	Trp	Tfg	Drg	-6.66	13.10μM	4
Glu	Pza	Fgl	Dlu	-6.65	13.25μM	4
Hgu	Ppg	Fg3	Leu	-6.65	13.30μM	5
Hgu	Ppg	Tfg	Dlu	-6.65	13.39μM	5
Hgu	Cha	Cpa	Dhr	-6.64	13.59μM	7
Hgu	Cha	Cpa	Val	-6.64	13.53μM	5
Asp	Ppg	Tfg	Leu	-6.64	13.51μM	4
Glu	Cha	Fg3	Dhr	-6.63	13.81μM	6
Glu	Pza	Pda	Asn	-6.63	13.72μM	2
Hgu	Cha	Fg2	Drg	-6.63	13.85μM	5
Hgu	Pza	Cha	Dlu	-6.63	13.80μM	6
Glu	Cha	Cha	Leu	-6.62	13.99μM	5
Hgu	Trp	Fg2	Leu	-6.61	14.38μM	6
Hgu	Trp	Fg3	Dlu	-6.61	14.29μM	4
Hgu	Trp	Tfg	Asn	-6.61	14.22μM	5
Asp	Pza	Cha	Drg	-6.61	14.36μM	4
Asp	Pza	Fg2	Dlu	-6.61	14.36μM	1
Glu	Trp	Fgl	Dhr	-6.6	14.52μM	5
Hgu	Pza	Cpa	Dhr	-6.6	14.60μM	6
Hgu	Cha	Tfg	Asn	-6.59	14.82μM	2
Hgu	Pza	Cha	Leu	-6.59	14.84μM	6
Hgu	Pza	Fg3	Val	-6.59	14.86μM	6
Asp	Pza	Cpa	Dlu	-6.59	14.85μM	2
Asp	Pza	Pda	Dhr	-6.58	14.99μM	5
Glu	Pza	Cha	Drg	-6.58	14.96μM	5
Glu	Pza	Tfg	Dlu	-6.58	14.92μM	3
Hgu	Pza	Tfg	Val	-6.58	14.99μM	2
Glu	Pza	Cpa	Val	-6.57	15.34μM	4
Glu	Pza	Fgl	Drg	-6.57	15.37μM	3
Glu	Trp	Fg3	Leu	-6.57	15.16μM	4
Hgu	Cha	Fg3	Dlu	-6.57	15.18μM	6
Hgu	Ppg	Fg2	Dlu	-6.57	15.22μM	4
Hgu	Ppg	Tfg	Drg	-6.57	15.23μM	5
Glu	Pza	Pda	Dhr	-6.56	15.61μM	7
Glu	Trp	Tfg	Drg	-6.56	15.57μM	5
Hgu	Pza	Cha	Dhr	-6.56	15.57μM	7
Hgu	Pza	Fg3	Leu	-6.56	15.58μM	4
Glu	Ppg	Cpa	Asn	-6.55	15.75μM	4
Hgu	Cha	Tfg	Val	-6.55	15.84μM	5
Hgu	Pza	Fg3	Dhr	-6.55	15.78μM	6
Hgu	Pza	Fgl	Drg	-6.55	15.75μM	5
Glu	Ppg	Fg3	Dhr	-6.54	16.14μM	7
Asp	Ppg	Cpa	Dhr	-6.54	16.09μM	6

Glu	Pza	Fg3	Dhr	-6.53	16.38μM	10
Hgu	Cha	Fgl	Leu	-6.53	16.39μM	5
Hgu	Pza	Tfg	Dlu	-6.53	16.30μM	4
Glu	Trp	Pda	Asn	-6.52	16.57μM	6
Glu	Pza	Cha	Asn	-6.51	16.88μM	3
Glu	Trp	Tfg	Dlu	-6.5	17.15μM	5
Hgu	Cha	Pda	Asn	-6.5	17.13μM	6
Glu	Trp	Fg3	Dlu	-6.49	17.53μM	5
Hgu	Cha	Fg2	Dhr	-6.49	17.38μM	7
Hgu	Cha	Tfg	Dhr	-6.49	17.40μM	6
Hgu	Trp	Cpa	Dhr	-6.49	17.57μM	7
Hgu	Pza	Cpa	Val	-6.48	17.69μM	6
Glu	Cha	Tfg	Dlu	-6.47	18.11μM	4
Glu	Ppg	Tfg	Dlu	-6.47	18.02μM	4
Hgu	Ppg	Fgl	Leu	-6.47	18.23μM	5
Hgu	Trp	Tfg	Dlu	-6.47	18.06μM	3
Hgu	Cha	Fg2	Leu	-6.46	18.53μM	7
Hgu	Ppg	Cpa	Val	-6.46	18.54μM	4
Hgu	Ppg	Pda	Drg	-6.46	18.25μM	8
Hgu	Trp	Cha	Drg	-6.46	18.53μM	8
Hgu	Trp	Fg2	Asn	-6.46	18.55μM	4
Hgu	Pza	Tfg	Asn	-6.45	18.73μM	3
Glu	Pza	Tfg	Dhr	-6.44	18.99μM	4
Glu	Pza	Fg2	Dlu	-6.43	19.34μM	3
Glu	Trp	Fg2	Leu	-6.43	19.50μM	8
Hgu	Pza	Fg2	Asn	-6.43	19.47μM	6
Hgu	Pza	Pda	Asn	-6.43	19.45μM	6
Glu	Cha	Cha	Val	-6.42	19.83μM	3
Glu	Trp	Cpa	Drg	-6.42	19.79μM	9
Hgu	Trp	Fg2	Dhr	-6.42	19.82μM	4
Glu	Cha	Tfg	Val	-6.41	20.04μM	4
Glu	Ppg	Pda	Leu	-6.41	20.14μM	4
Glu	Trp	Cpa	Dlu	-6.41	20.07μM	8
Asp	Ppg	Pda	Drg	-6.41	20.06μM	7
Asp	Cha	Cpa	Dhr	-6.41	19.89μM	7
Hgu	Cha	Cpa	Dlu	-6.4	20.28μM	5
Hgu	Ppg	Cha	Asn	-6.4	20.20μM	6
Hgu	Trp	Fgl	Asn	-6.4	20.31μM	6
Glu	Cha	Pda	Dlu	-6.39	20.68μM	4
Glu	Trp	Tfg	Val	-6.39	20.55μM	4
Glu	Ppg	Fg2	Leu	-6.37	21.28μM	5
Hgu	Trp	Fgl	Leu	-6.35	22.25μM	5
Hgu	Cha	Fgl	Dhr	-6.34	22.56μM	6
Hgu	Pza	Tfg	Dhr	-6.34	22.43μM	7
Hgu	Trp	Pda	Val	-6.32	23.12μM	5
Glu	Trp	Cha	Asn	-6.31	23.83μM	4
Hgu	Ppg	Cha	Dlu	-6.31	23.56μM	7
Hgu	Ppg	Tfg	Val	-6.31	23.63μM	4

Hgu	Trp	Cha	Val	-6.31	23.68μM	6
Asp	Trp	Cha	Dhr	-6.28	25.00μM	4
Glu	Ppg	Cha	Drg	-6.27	25.54μM	6
Hgu	Trp	Cpa	Val	-6.27	25.39μM	7
Hgu	Cha	Fg3	Leu	-6.25	26.41μM	6
Hgu	Ppg	Cpa	Dhr	-6.25	26.03μM	10
Asp	Ppg	Cpa	Leu	-6.23	26.95μM	4
Glu	Cha	Fg3	Leu	-6.22	27.48μM	6
Glu	Ppg	Cha	Dhr	-6.22	27.57μM	9
Glu	Ppg	Cpa	Dlu	-6.22	27.80μM	6
Hgu	Trp	Cpa	Drg	-6.21	28.06μM	10
Glu	Pza	Tfg	Val	-6.2	28.46μM	3
Hgu	Trp	Tfg	Dhr	-6.19	28.87μM	5
Glu	Trp	Tfg	Leu	-6.17	29.96μM	5
Hgu	Ppg	Fg2	Asn	-6.17	29.79μM	5
Hgu	Pza	Fg2	Val	-6.15	30.87μM	3
Hgu	Trp	Cha	Leu	-6.15	31.07μM	6
Glu	Ppg	Tfg	Leu	-6.14	31.55μM	5
Hgu	Pza	Fgl	Dlu	-6.14	31.35μM	4
Glu	Pza	Cpa	Drg	-6.13	32.20μM	4
Hgu	Trp	Pda	Dlu	-6.13	31.91μM	7
Glu	Trp	Cha	Leu	-6.12	32.42μM	8
Glu	Ppg	Cpa	Leu	-6.11	33.03μM	3
Glu	Pza	Pda	Drg	-6.11	33.02μM	5
Glu	Pza	Cha	Dlu	-6.1	33.51μM	7
Hgu	Ppg	Cpa	Asn	-6.1	33.99μM	4
Asp	Pza	Tfg	Dlu	-6.09	34.32μM	1
Hgu	Trp	Cha	Dlu	-6.09	34.42μM	6
Glu	Pza	Tfg	Leu	-6.08	34.99μM	3
Hgu	Ppg	Cha	Val	-6.08	34.97μM	8
Hgu	Trp	Cpa	Dlu	-6.08	34.77μM	6
Hgu	Cha	Cha	Drg	-6.07	35.59μM	5
Hgu	Trp	Cpa	Leu	-6.07	35.40μM	8
Hgu	Trp	Tfg	Leu	-6.06	36.16μM	6
Hgu	Cha	Cha	Asn	-6.05	36.57μM	4
Hgu	Ppg	Tfg	Asn	-6.02	38.82μM	5
Glu	Cha	Cha	Asn	-6.01	39.51μM	6
Hgu	Ppg	Pda	Val	-6	39.81μM	8
Hgu	Cha	Tfg	Leu	-5.99	40.66μM	6
Glu	Trp	Cha	Drg	-5.98	41.11μM	6
Glu	Pza	Cpa	Dhr	-5.97	42.12μM	6
Hgu	Ppg	Fg2	Val	-5.96	42.86μM	6
Hgu	Trp	Pda	Drg	-5.96	42.79μM	7
Glu	Cha	Cha	Dlu	-5.93	45.13μM	5
Glu	Ppg	Cpa	Drg	-5.92	45.92μM	9
Glu	Trp	Pda	Dlu	-5.85	51.47μM	9
Glu	Pza	Pda	Dlu	-5.83	53.43μM	7
Glu	Cha	Cha	Drg	-5.82	54.46μM	6

Hgu	Cha	Pda	Dlu	-5.82	54.64μM	7
Hgu	Ppg	Cpa	Dlu	-5.81	55.48μM	8
Hgu	Ppg	Cha	Leu	-5.79	56.55μM	5
Hgu	Ppg	Fg2	Leu	-5.79	56.68μM	6
Hgu	Trp	Cha	Asn	-5.77	59.26μM	9
Hgu	Pza	Fg2	Leu	-5.75	61.13μM	6
Glu	Trp	Cpa	Dhr	-5.72	64.11μM	5
Hgu	Ppg	Fgl	Dhr	-5.72	64.27μM	8
Hgu	Ppg	Pda	Leu	-5.71	65.03μM	8
Hgu	Pza	Cpa	Drg	-5.68	69.18μM	7
Hgu	Ppg	Cpa	Leu	-5.67	69.90μM	4
Hgu	Cha	Cha	Dlu	-5.63	74.14μM	7
Glu	Trp	Tfg	Dhr	-5.62	75.44μM	6
Hgu	Pza	Pda	Dlu	-5.62	76.37μM	8
Hgu	Ppg	Pda	Dhr	-5.58	81.86μM	9
Hgu	Ppg	Fg2	Dhr	-5.57	81.94μM	10
Hgu	Cha	Cha	Dhr	-5.49	94.31μM	7
Hgu	Ppg	Fg3	Dhr	-5.43	104.30μM	8
Hgu	Trp	Fg3	Leu	-5.43	104.57μM	4
Hgu	Trp	Cha	Dhr	-5.38	114.42μM	8
Glu	Cha	Cpa	Dhr	-5.36	118.73μM	7
Hgu	Ppg	Tfg	Leu	-5.32	126.43μM	4
Hgu	Pza	Pda	Dhr	-5.29	132.61μM	9
Glu	Trp	Cha	Dhr	-5.26	138.36μM	7
Hgu	Pza	Pda	Drg	-5.24	145.24μM	8
Hgu	Cha	Pda	Dhr	-5.22	147.94μM	7
Hgu	Ppg	Cpa	Drg	-5.22	148.37μM	10
Hgu	Ppg	Pda	Dlu	-5.13	174.14μM	8
Hgu	Ppg	Cha	Dhr	-4.99	218.49μM	8
Glu	Cha	Cha	Dhr	-4.76	323.46μM	4
Hgu	Trp	Pda	Dhr	-4.64	397.73μM	8

5.4.6 Benchmarking against related structures

The library was built and optimized for the best interaction mimicking the *Blys*/*BaffR* complex (as given in PDB structure 1OSG). Several related structures exist in the PDB: *Blys*/*Bcma* complex (PDB 1OQD), another crystal structure for *Blys*/*BaffR* complex with minor structural changes in *Blys* (PDB 1OQE), *April*/*Bcma* complex (PDB 1XU2), *April*/*Taci* complex (PDB 1XU1), and the apo structure of *Blys* (PDB

1KXG). We now test the performance of the top 100 compounds from the ligand library against each of these structures.

The top 100 ligands from the library were docked with Autodock [71]. We then assess the performance of these ligands using several criteria including:

1. the minimum, maximum, average, and standard deviation of the predicted energies of interaction
2. the number of clusters per docking generated
3. the percentage of the compounds mimicking the original pose of the ligand in each structure

Table 5-9: The docking of the top 100 ligands from cyclic hexapeptide library (Table 5-8) onto 6 structures. *Blys*/*Bcma* complex (PDB 1OQD), *Blys*/*BaffR* complex (PDB 1OQE), *April*/*Bcma* complex (PDB 1XU2), *April*/*Taci* complex (PDB 1XU1), the apo structure of *Blys* (PDB 1KXG), and the *Blys*/*BaffR* complex used in this chapter (PDB 1OSG). The correct pose is defined as any pose mimicking the original ligand binding mode. The "average clusters per compound" denotes average of the number of clusters in which the 10 docking experiments per molecule can be grouped (the less the clusters, the better the confidence).

PDB	Minimum energy (kcal/mol)	Maximum energy (kcal/mol)	Average energy (kcal/mol)	Stdev of energy (kcal/mol)	% structures in correct Pose	Average clusters per compound
1OQD	-8.54	-5.74	-7.14	0.609	66	6.06
1OQE	-9.04	-5.53	-7.34	0.57	83	5.38
1XU2	-9.16	-4.1	-6.8	1.11	22	5.98
1XU1	-8.55	-4.83	-6.55	0.79	31	5.88
1KXG	-9.18	-5.84	-7.31	0.603	34	6.08
1OSG	**-8.9**	**-7.47**	**-7.81**	**0.28**	**95**	**3.66**

The docking of the top 100 ligands was assessed based on the attributes discussed in Table 5-9. Some compounds showed a higher affinity towards the related structures tested (shown with the minimum energy in some of these experiments being more negative than the minimum energy in the *Blys*/*BaffR* complex). However, the overall performance of these ligands was best with the original *Blys*/*BaffR* complex (PDB 1OSG). When docked to 1OSG, the ligands exhibited the lowest average energy, the smallest standard deviation, the lowest maximum energy, and the highest percentage of ligand orientations mimicking the original *BaffR* binding mode. We also studied the average number of clusters per docked ligand for each structure. Autodock generates the best 10 poses per ligand and ranks them based on binding energy. These poses are then clustered based on rmsd. In theory, if all the top 10 poses belong to the same cluster, then we have a higher confidence that the virtual docking screen is accurate and will mimic the reality. The more clusters we have, the less our confidence in the docking result. The ligands showed the lowest average clusters per docking in case of 1OSG, indicating a more reliable prediction of the correct docking position.

The related structures included 3 *Blys* structures (1OQD, 1OQE, and 1KXG) and 2 *April* structures (1XU1 and 1XU2). The ligands showed a more negative average energy with the *Blys* structures when compared with *April*. With the exception of 1KXG which is the *Blys* Apo structure, the ligands had a high percentage of correct poses (poses that mimic the original ligand binding mode) when docked on these *Blys* structures (66% and 83%). The percentage of correct poses with the *April* structures is low (22% and 31%). The low percentage of correct pose on the *Blys* apo

structure (34%) is expected as the pocket undergoes structural changes upon binding. This makes it hard to mimic the *BaffR* binding pose with the Apo structure.

We conclude that the created ligand library has a better binding affinity to *Blys* structures and very successful at mimicking the *Blys*/*BaffR* interaction. Some of these ligands do bind tightly to *April*, and this is not surprising as *April* and *Blys* bind common ligands. However the ligands in the created library show a preference to bind to 1OSG and analyses of the docking experiments show a high confidence in the docking results with 1OSG. Synthesis of the molecules has been designed by Auer et al. The next step is to synthesize these molecules and try them *in vitro*.

6 Screening for Inhibitors of Protein - Peptide Interaction

6.1 Introduction

Protein-protein interactions are important targets for the development of therapeutics [252-257]. Often, protein-protein interaction is determined by a short contiguous stretch of aminoacids that acts as the recognition and binding signal. For example, MDM2/p53 [258, 259], PCNA/p21 [260-262] and eIF-4E/eIF-BP1 [263] form such complexes. There is a growing interest in such linear peptide motifs [264], as highlighted by resources like the Eukaryotik Linear Motif Server (ELM) [265]. Small-molecule inhibitors [252, 253, 255, 256, 266-268] targeted at protein-protein interactions have been described [258, 269-278]. These include nutlin (inhibits MDM2 and p53 interaction [258, 259]), compounds which inhibit members of the apoptosis regulating Bcl-2 family of proteins [270, 274, 275, 277], and mimics blocking the interaction of Smac (second mitochondria-derived activator of caspases) and X chromosome-encoded inhibitor-of-apoptosis protein (XIAP) [269].

The growing awareness of the biological and medical importance of protein-peptide interactions as around 40% of signal transduction events are dependent on such interactions [264, 279]. This has led to research and analysis of peptide structural motifs [280] and the development of searchable databases like PepX [99]. ProPep [103] is an inhouse repository for X-ray structural data, used to study trends, motifs, residue pairing frequencies, and amino acid enrichment propensities in protein-peptide interaction. This chapter discusses updating and automating the ProPep

database to the November 2008 version of the PDB. As an example of how the database can be used, a study of the interactions made by the LxxL α helical binding motif is presented. The incorporation of virtual screening programs like UFSRAT [98] and AutoDock [71] into this process leads to the development of a list of possible peptidomimetic small molecules for a subclass of this LxxL α helical motif.

6.2 The ProPep Database

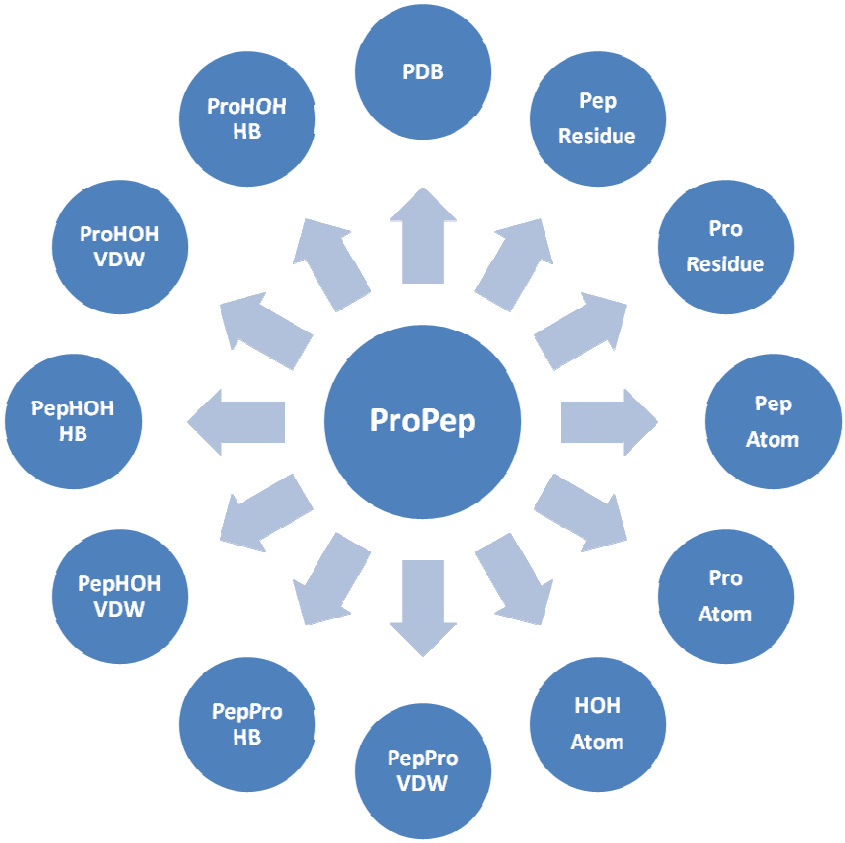

Figure 6-1: The ProPep dataset consists of 12 tables. PDB holds all information about the PDB structures in the database and what chains of these structures are used for the protein and peptide sides of the interface. ProResidue and PepResidue hold information about the residues of the protein and peptide chains. ProAtom, PepAtom, and HOHAtom hold information about the atoms of the protein and peptide chains and the water molecules in the dataset as well. PepProVDW and PepProHB hold information about the VDW and hydrogen bonds between the peptides and proteins in the interface. PepHOHVDW and PepHOHHB hold information about the VDW and hydrogen bonds between the peptide atoms and water molecules. Finally, ProHOHVDW and ProHOHHB hold information about the VDW and hydrogen bonds between the protein atoms and water molecules.

The ProPep database [103] is designed to study the binding of short contiguous peptides to proteins (Figure 6-1). This type of binding is a good target for inhibiting protein-protein interaction. This repository conatins a representative dataset of all protein-peptide interactions in the PDB [17]. The February 2007 version of the PDB is represented by a dataset of 274 Protein-peptide complexes. This dataset represents all 3.0 Å resolution (or better) X-ray structures of complexes made up of 2 to 6 chains (excluding nucleic acid chains), of which one is between 3 and 50 residues in length (the peptide) and one between 50 and 600 residues (the protein). All the structures in the dataset share less than 90% sequence identity between each other. Measures are taken to ensure the protein and the peptide are interacting and that membrane proteins are excluded [103]. This database lists and studies all VDW and hydrogen bond interactions between the protein and the peptide, the protein and water molecules, and the peptide and water molecules.

6.3 Updating the ProPep database

A newer version of the ProPep database has been created to represent all the structures existent in the PDB as of November 2008 (54077 PDB structures). The new database, ProPep08, comprised 481 protein-peptide interaction complexes. This database is stored on ‘ocycfs’ and can be accessed via:

mysql -h ocycfs -u <username> -p<password> ProPep08

Compiling the ProPep database is divided into 2 stages. First, all structures are checked for eligibility and sequence identity to create the smallest possible representative dataset of all protein-peptide interactions in the PDB. Then, these

chosen structures are processed and analyzed and the data is all recorded in the ProPep database. This is a very lengthy process and is performed by a group of scripts that have been designed specifically for this purpose [103]. These scripts are located on the scibs file system in the directory:

'/usr/people/wissam/Simon/perlScripts/wissam'.

An automation script (Appendix Programming Code 12-1) is created (scripts.c). This script facilitates the creation of the entire dataset without human intervention. A readme file is present in this directory, indicating how to compile hbplus [281], naccess [282], and torsion [283]. A programming bug in newrestoatom.pl was found and fixed. The program is now capable of dealing with with protein atoms numbered "0000". Moreover, a similar script (additionalUpdates.c) is created to update the VDW, Node, HB, Prob (problems), and Contact columns in the database (Appendix Programming Code 12-2). The calculation of the data needed to produce the analysis graphs introduced by [103] is also automated (Appendix Programming Code 12-3).

6.4 Changes to the database upon the update

6.4.1 Database Size

The number of complexes constituting the database increased from 274 to 481 (76% increase). This resulted in a similar increase in the size of the 12 tables constituting the database (Table 6-1). All tables increased consistently (around 72% increase), except for ProHOHHB and PepHOHHB which record the hydrogen bond interaction between protein atoms and water, and peptide atoms and water respectively. ProHOHHB and PepHOHHB are small tables (Table 6-1) which may account for the inconsistent change in their sizes.

Table 6-1: The ProPep database tables and their functions.

Table Name	Data Summary	Number of entries
PDB	Complexes constituting the database	481
ProResidue	All protein residues in the database	12225
ProAtom	All protein atoms in the database	102334
PepResidue	All peptide residues in the database	6015
PepAtom	All peptide atoms in the database	48232
HOHAtom	All water molecules in the database	7340
PepProVDW	Protein-peptide VDW interactions	323838
PepProHB	Protein – Peptide hydrogen bond interactions	4000
ProHOHVDW	Protein – Water VDW interactions	11164
ProHOHHB	Protein – Water HB interactions	1770
PepHOHVDW	Peptide – Water VDW interactions	18050
PepHOHHB	Peptide – Water HB interactions	4616

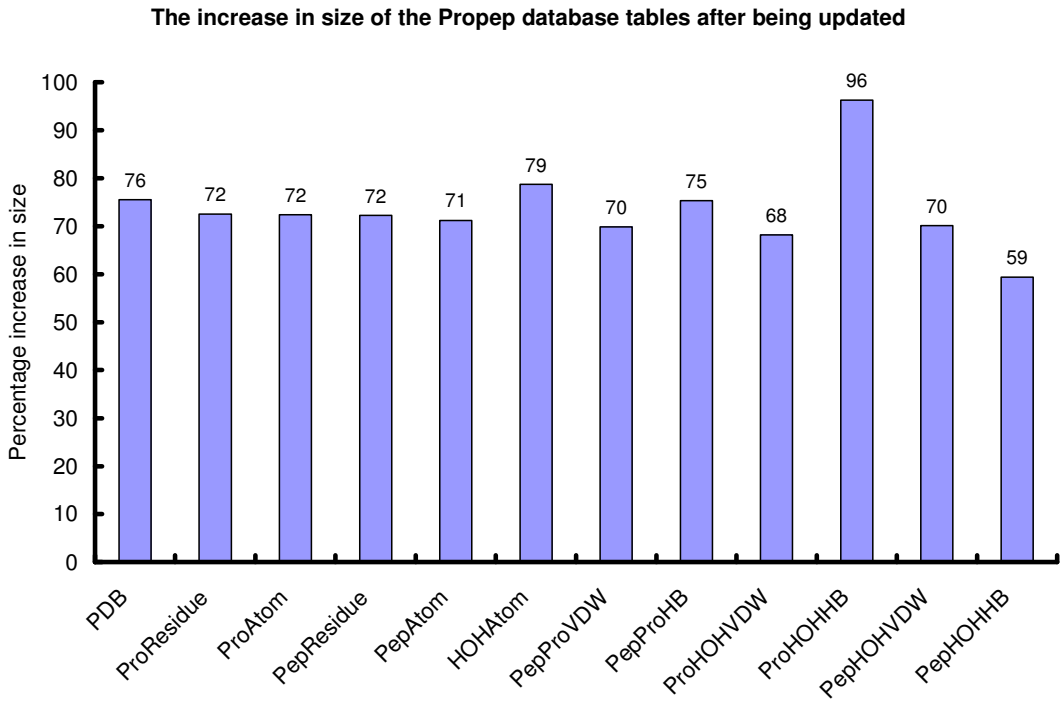

Figure 6-2: The increase in size of the database tables upon updating the ProPep database to the November 2008 version of the PDB.

6.4.2 Relative Abundance of Amino acids and Residue Pairing Preference

$$Enrichment = \%\ composition\ in\ ProPep - \%\ composition\ in\ SwissProt$$

Equation 6-1: Calculation of the enrichment of amino acid composition in protein – peptide interaction sites. Three enrichment values are calculated for each of the 20 amino acid types: the first for enrichment on the peptide side of the interface, the second for enrichment on the protein side of the interface, and the third for the entire protein – peptide interface.

The ProPep database allows the study of the relative abundance of amino acids at the protein-peptide interface. Each amino acid has an expected frequency of occurrence according to the Swissprot database. The observed frequency of occurrence for each amino acid on the peptide and protein side of the interface is compared with its expected frequency, yielding an enrichment value (Equation 6-1). A positive enrichment value indicates an increased expression of a certain amino acid. The relative abundance frequencies for the 20 amino acids in the earlier and updated version of ProPep are shown (Figure 6-3). Pro and Leu occur more frequently on the peptide side of the interface and less frequently on the protein side while the aromatic residues His, Phe, Trp, and Tyr occur more frequently in ProPep than in Swissprot [103]. Only subtle differences exist between the enrichment values of amino acids in the February 2007 and the November 2008 versions of ProPep. The absolute value of the change of enrichment per aminoacid on the peptide side of the interface (between the 2007 and 2008 versions of ProPep) has an average of 0.41% and a standard deviation of 0.20% with a maximum change observed for Asn at 0.76%. As for the protein side of the interface, the change in enrichment has an average of 0.23% and a standard deviation of 0.16%, with a maximum change observed for Leu at 0.71%. For the interface as a whole, the change in enrichment

has an average of 0.22% with a standard deviation of 0.11%, and a maximum value observed for Gly at 0.37% (). The absence of major differences between the two versions of ProPep (Table 6-2) validates the conclusions drawn from the relative abundance of amino acids.

Table 6-2: The change of amino acid enrichment between the Nov 2008 and Feb 2007 versions of ProPep. A positive value shows an increase in 2008. Enrichments are calculated as percentage differences between the composition of protein-peptide interfaces in the ProPep database and Swissprot.

Residue	Peptide Side of the Interface (%)	Protein Side of the Interface (%)	Entire Interface (%)
Ala	-0.59	-0.20	-0.32
Leu	-0.39	0.71	0.36
Pro	0.30	0.31	0.33
Val	-0.20	0.22	0.08
Ser	0.67	-0.15	0.13
Gly	-0.59	-0.25	-0.37
Glu	-0.13	-0.28	-0.24
Lys	-0.28	-0.28	-0.28
Ile	0.43	-0.47	-0.16
Thr	0.55	0.04	0.21
Cys	0.18	0.18	0.18
Asp	0.50	0.21	0.30
Gln	-0.48	0.12	-0.08
Met	-0.21	-0.32	-0.29
Asn	0.76	0.16	0.36
His	-0.26	-0.15	-0.19
Arg	0.61	-0.15	0.10
Phe	-0.65	-0.03	-0.24
Trp	0.11	-0.01	0.02
Tyr	-0.34	0.36	0.11
Average	**0.41**	**0.23**	**0.22**
StDev	**0.20**	**0.16**	**0.11**

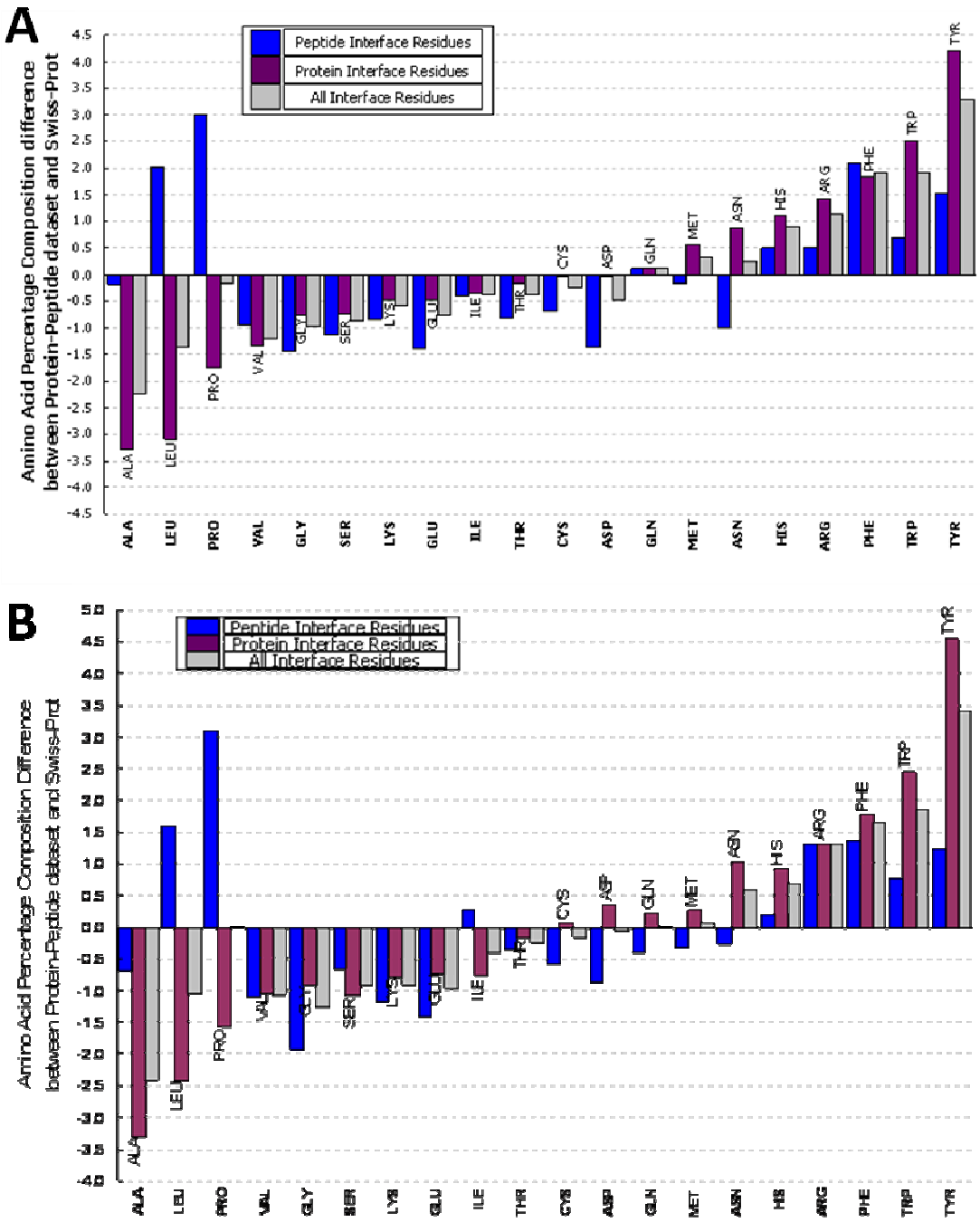

Figure 6-3: Calculation of the enrichment of the occurrence of Amino Acids (Equation 6-1) at different sides of the protein peptide interface according to the February 2007 (A) and November 2008 (B) versions of ProPep. Results show a Preference for Leu and Pro to overexpress on the peptide side of the interface as well as a general over-occurrence of the aromatic residues His, Phe, Trp, and Tyr.

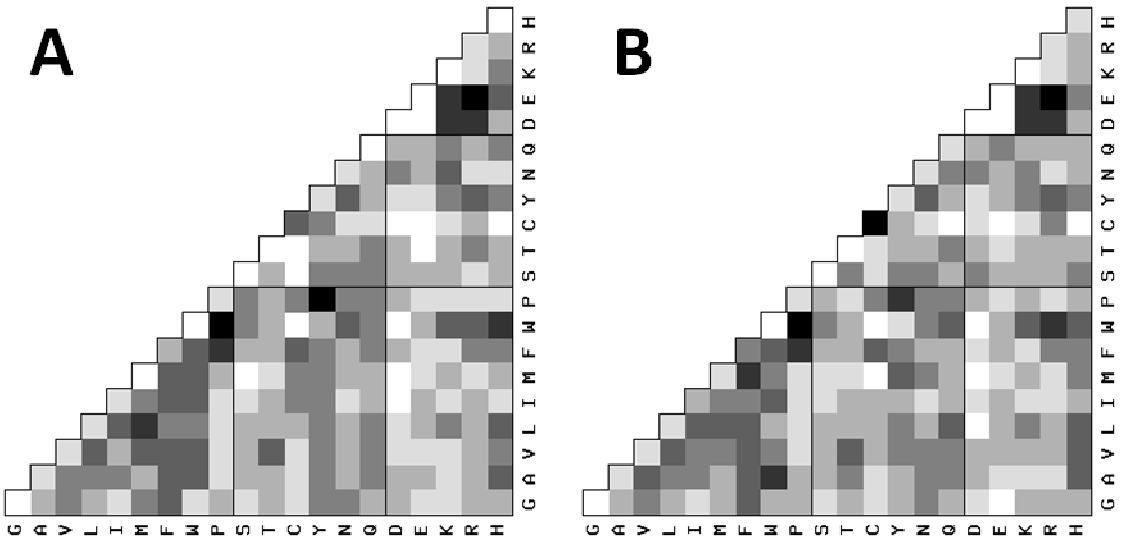

Figure 6-4: Residue Pairing Preference Diagrams in the February 2007 (A) and the November 2008 (B) versions of ProPep. Both graphs show high pairing preference between Trp and His, Glu and Arg, Pro and Tyr, Pro and Trp, and Cys and Cys residues.

$$RPP(i,j) = \frac{Probability\ of\ finding\ pair\ (i,j)}{Probability\ of\ finding\ AA\ i\ \times Probability\ of\ finding\ AA\ j}$$

$$RPP(i,j) = \frac{\frac{NumPairs(i,j)}{Total\ NumPairs}}{\frac{NumAA(i)}{TotalNumAA} \times \frac{NumAA(j)}{TotalNumAA}}$$

Equation 6-2: The calculation of the residue pairing frequencies of 2 amino acids i and j, where NumPairs (i, j) is the number of observed pairs (i, j) in the dataset, and NumAA(i) is the number of occurrences of amino acid i in the dataset. A pair is defined as 2 aminoacids on opposite sides of the protein-peptide interface that are in contact with each other.

The ProPep database defines the Residue Pairing Preference (RPP, Equation 6-2) as a normalized probability score of a particular pair of residues interacting across the protein-peptide interface [103]. The study of residue pairing in Pro-Pep showed high RPP values for salt bridges: Lys-Asp, Lys-Glu, Arg-Asp, Arg-Glu (3.95, 4.02, 4.56

and 5.31 respectively). These interactions have been documented in the literature [284-286]. Pro pairs frequently with Tyr (4.28) and Trp (5.51), possibly due to Pro's tendency to interact with aromatic residues in many different modes [282, 286]. Updating the ProPep database produced no significant changes in the RPP of the 20 amino acids (Figure 6-4).

6.4.3 The participation of protein atoms in protein – peptide interaction

The ProPep database defines every protein atom as a Node, based on the position it occupies, and not the atom type (Figure 6-5). The main chain nitrogen, carbonyl, and oxygen atoms occupy positions M2, M1, and M3 respectively, while the side chain atom positions are named according to their PDB atom order (α, β, gamma, delta, epsilon, zeta, and eta). The reduction of accessible surface area (RASA, Equation 6-3) of an atom is defined as the difference between the accessible surface area of a certain atom before and after binding of the protein to the peptide [103]. The pictogram (Figure 6-5 and Figure 6-6) illustrates the RASA of every possible protein node on a green to red scale. A higher RASA value indicates a more involved role in interaction. It is therefore not surprising that the further the atom is from the mainchain, the more involved it is in protein peptide interaction, with the exception of the main chain oxygen atom which plays an important role in hydrogen bond interactions across the protein-peptide interface [103]. Updating the ProPep database did not change the output of this pictogram significantly (Figure 6-5 and Figure 6-6).

$$RASA(i) = ASA\ (i)\ before\ binding - ASA\ (i)\ after\ binding$$

where ASA (i) is the accessible surface area of a amino acid i
and RASA (i) is the reduction in accessible surface area of amino acid i

Equation 6-3: Calculation of the reduction in accessible surface area of amino acids

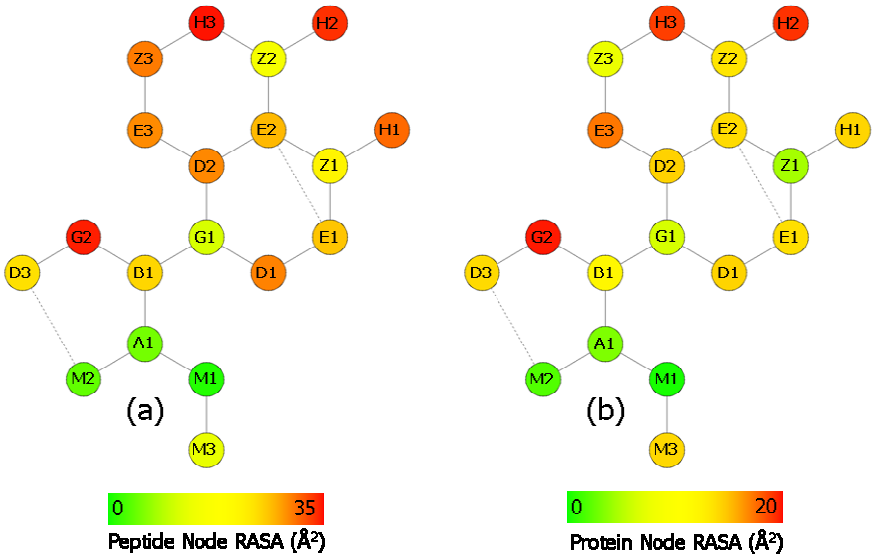

Figure 6-5: The Amino acid Pictogram is used to show the Reduction of Accessible Surface Area (RASA) on different amino acid nodes on the Protein and Peptide side of the interface in the entries of the February 2007 version of the ProPep. In general, the RASA increases the further the atom is from the mainchain. The main chain oxygen displays a high RASA because of its role in hydrogen bond interactions across the protein – peptide interface.

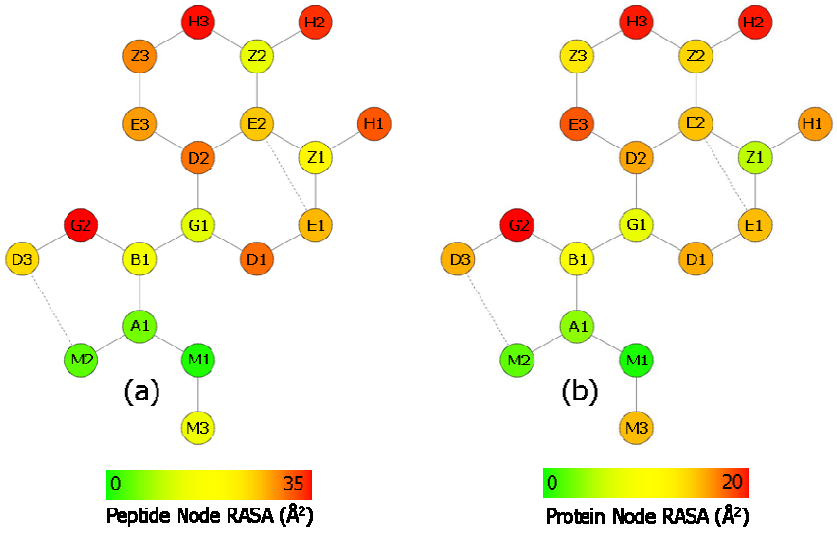

Figure 6-6: The Amino acid Pictogram is used to show the Reduction of Accessible Surface Area (RASA) on different amino acid nodes on the Protein and Peptide side of the interface in the entries of the November 2008 version of the ProPep. In general, the RASA increases the further the atom is from the mainchain. The main chain oxygen displays a high RASA because of its role in hydrogen bond interactions across the protein – peptide interface.

6.5 The LxxL Interaction Motif

Molecular function relies on the interaction between short segments of conserved sequence and/or structural elements that are vital to the fold and function of a protein [287-289]. The importance of these short segments is evident in the fact that homologous proteins and proteins with similar folds can perform distinct biochemical functions [290] while proteins with different folds can perform the same function with the same set of residues and a similar mechanism [291]. These segments, or motifs, are widely used as markers for protein function and are useful targets for inhibiting a certain class of molecular interaction.

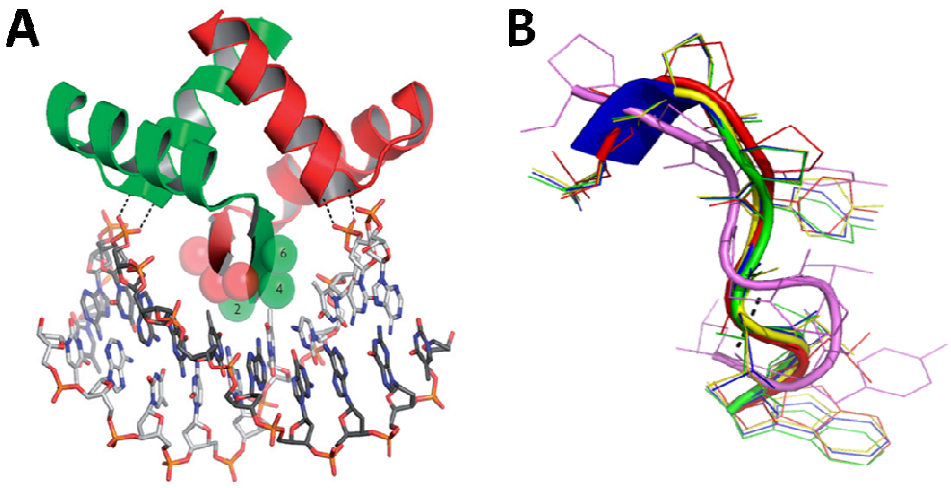

Figure 6-7: Motifs in protein interaction: (A) shows a model of the Ribbon-Helix-Helix dimer interacting with a DNA molecule, Figure taken from Schreiter et al, (2007) [292]. (B) shows the structures of the heme-binding loop in heme-copper oxidases [W/Y]xxYPPL, Figure taken from Marsico et al, (2010) [293].

Many motif mining programs are currently available like Rasmot-3D [294], MultiBind [295], MegaMotifBase [289], PAR-3D [296], Superimpose [297], and the SPASM server [298]. Many examples of motifs are discussed in the literature, like

the ribbon-helix-helix (RHH) motif that is a conserved three dimensional motif used to bind to DNA [292] (Figure 6-7 A). Another example is the [W/Y]xxYPPL heme-binding loop in heme-copper oxidases responsible for the catalysis of the reduction of molecular oxygen to water, a reaction used to produce chemical energy needed to transfer protons across the cell membrane and generate an electro chemical proton gradient [293] (Figure 6-7 B).

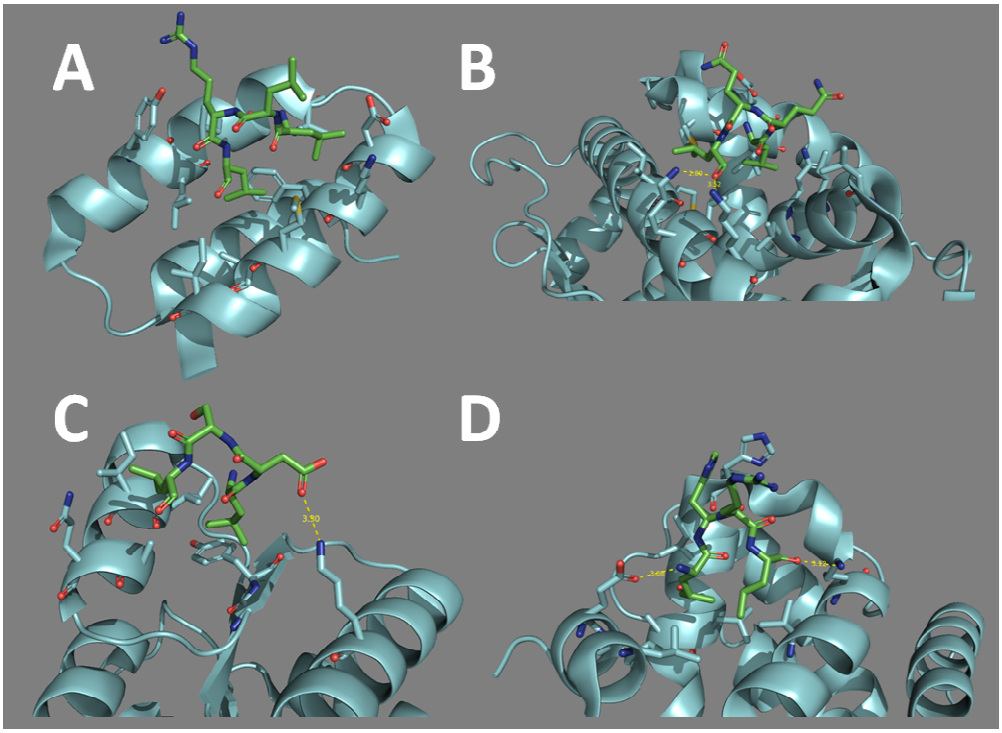

Figure 6-8: The interaction of the LxxL α helical motif with various structures in the dataset. (A) shows a viral protein, paramyxoviral polymerase (pdb 1T6O). (B) shows a transferase, Mineralocorticoid Receptor (pdb 2A3I). (C) shows a signalling protein, chemotaxis protein cheY (pdb 2FMK). (D) shows a transcription protein, the farnesoid X receptor (pdb 3BEJ).

The LxxL α-helical is another motif and has many documented roles in various biological pathways (Figure 6-8). The biological functions played by this motif include a steroid-binding role in the large family of human and yeast oxysterol binding related proteins [299], a co-repressor role of the transcriptional and growth

arrest activities of the CCAAT/enhancer-binding protein α [300], a coactivator role of the androgen receptor [301], and Alix-mediated budding of HIV [302, 303].

Being an α-helical motif, the structure of the main chain of this motif is conserved, with the only alterations happening at the side chains of the XX amino acids in the motif. The role of the XX amino acids in ProPep is discussed later in this chapter (Table 6-3 and Table 6-4). This motif is a strong candidate for drug discovery, due to its numerous applications and conserved structural attributes. We use the ProPep database to search for the occurrence of this motif, its roles, variations, and pave the way towards ligand based virtual screening in a search for LxxL-mimic compounds.

The ProPep database is designed to enable fast motif searches. Searching for the LxxL α helix motif yielded 42 protein-peptide interfaces (8.7%), distributed on 18 protein classes (Table 6-3). Four of these classes stand out for binding an LxxL motif frequently in ProPep (Figure 6-9). Of these 4 classes, “transcription proteins” is the largest, with 17 structures (16 nuclear receptors and 1 nuclear coactivator, 1OJ5) containing the LxxL motif out of the 39 transcription-class structures in the database (43%).

Table 6-3: The occurrence of the LxxLx α helical motif in the ProPep dataset. The dominance of Leu in the last residue (number 5) is obvious at 60%.

PDB	Protein Classification	Motif (LxxLx)
1LQV	BLOOD CLOTTING	LxxLR
2VZG	CELL ADHESION	LSELD
2UWJ	ChaPERONE	LRDLM
1YCQ	COMPLEX (ONCOGENE ProTEIN/PEPTIDE)	LWKLL
1UHL	DNA BINDING ProTEIN	LHRLL
2HC4	GENE REGULATION	LHRLL
2P1T	HORMONE RECEPTOR	LHRLL
2DOH	HYDROLASE	LQRLK
2BEC	MetAL BINDING ProTEIN/TRANSPORT ProTEIN	LDHLL
3BQD	ProTEIN BINDING	LQQLL
2O02	ProTEIN BINDING/TOXIN	LDALD
1J2J	ProTEIN TRANSPORT	LARLL
3CM8	RNA BINDING ProTEIN/TRANSFERASE	LLFLK
2OM2	SIGNALING ProTEIN	LVELL
2FMK	SIGNALING ProTEIN	LDSLG
1FQJ	SIGNALING ProTEIN	LHELA
2C23	SIGNALING ProTEIN/COMPLEX	LDALD
2F31	STRUCTURAL ProTEIN	LEALQ
1NQ7	TRANSCRIPTION	LHRLL
1PZL	TRANSCRIPTION	LHRLL
2Q3Y	TRANSCRIPTION	LYALL
2QZO	TRANSCRIPTION	LHRLL
1YMT	TRANSCRIPTION	LYALL
2P54	TRANSCRIPTION	LHRLL
2ZMI	TRANSCRIPTION	LMNLL
1YYE	TRANSCRIPTION	LVQLL
1ZH7	TRANSCRIPTION	LYTLL
3D24	TRANSCRIPTION	LKYLT
1NRL	TRANSCRIPTION	LHRLL
2GPO	TRANSCRIPTION	LLHLL
1YUC	TRANSCRIPTION REGULATION	LYALL
3BEJ	TRANSCRIPTION REGULATOR	LHRLL
1ZOQ	TRANSCRIPTION/TRANSFERASE	LQDLL
1XIU	TRANSCRIPTION/TRANSFERASE	LHRLL
1OJ5	TRANSCRIPTIONAL COACTIVATOR	LTKLL
1OW6	TRANSFERASE	LDELM
2A3I	TRANSFERASE	LQQLL
2VGO	TRANSFERASE	LEELF
2QN6	TRANSLATION	LDRLY
1T6O	VIRAL ProTEIN	LLRLQ
1SVF	VIRAL ProTEIN	LQHLA
1FAV	VIRAL ProTEIN	LLELD

Table 6-4: The occurrence of the LxxLx α helical motif in the ProPep database. The sequences are aligned to show the LxxL motif highlighted in Yellow. The LxxLL motifs (25 out of 42 structures) are further highlighted with green. Ten nuclear receptor structures discussed later in this chapter are highlighted in bold font. Their extended motif shared by these receptors (HKILHRLLQ) is highlighted further in grey.

PDB	Peptide Sequence
1J2J	----------NVIFEDEEKSKMLARLLKSSHPEDLRAANKLIKEMVQEDQKRMEK----------
2C23	----------GLLDALDLASK----------
2O02	----------GHGQGLLDALDLAS----------
1OW6	----------ATRELDELMASLS----------
2BEC	----------VDLLAVKKKQETKRSINEEIHTQFLDHLLTGIEDICGHYGHHH----------
2QN6	----------SSEKEYVEMLDRLysKLP----------
2FMK	----------ASQDQVDDLLDSLGF----------
2F31	----------DETGVMDSLLEALQSGAAFR----------
2VGO	PIPAWASGNLLTQAIRQQYYKPIDVDRMYGTIDSPKLEELFNKS----------
1FQJ	----GVQGFGDDIPGMEGLGTDITVICPWEAFNHLELHELAQYGII----------
1NQ7	----------HKILHRLLQE----------
1NRL	----------CPSSHSSLTERHKILHRLLQEGSPS----------
1PZL	----------HKILHRLLQEGSPS----------
1UHL	----------HKILHRLLQD----------
1XIU	----------RHKILHRLLQEGSPS----------
2HC4	----------RHKILHRLLQEGSPS----------
2P1T	----------KHKILHRLLQDSS----------
2P54	----------ARHKILHRLLQE----------
2QZO	----------KHKILHRLLQDSS----------
3BEJ	----------CPSSHSSLTERHKILHRLLQEGSPS----------
3D24	----------QQQKPQRRPCSELLKYLTTNDD----------
1FAV	---------XEXNNYTSLIHSLIEESQNQQEKNEQELLELDK----------
3CM8	----------GPLGSMDVNPTLLFLKVPAQNAISTTFPYT----------
2GPO	----------LERNNIKQAANNSLLLHLLKSQTIP----------
1T6O	----------QDSRRSADALLRLQAMAGIS----------
2ZMI	----------KNHPMLMNLLKDN----------
1ZOQ	----------SALQDLLRTLKSPSSPQQQQQVLNILKSNPQLMAAFIKQRTAKYVAN
1SVF	----------QILSIDPLDISQNLAAVNKSLSDALQHLAQSDTYLSAI----------
2A3I	----------QQKSLLQQLLTE----------
3BQD	----------AQQKSLLQQLLTE----------
2DOH	----------VEKLTADAELQRLKNERHEEAELERLLSEY----------
2UWJ	----------MHKINKWSVIYNINSTVTRALRDLMQGILQKI----------
2VZG	----------NLSELDRLLLELNAVQHNPP----------
1OJ5	----------LPPTEQDLTKLLLE----------
2OM2	----------DIEGLVELLNRVQSSGAHDQRGLLRKEDLVLPEFLQ----------
1YYE	----------SGSHKLVQLLTTT----------
1YCQ	----------PLSQETFSDLWKLLPEN----------
1LQV	----------ANSFLxxLRHSSLXRXCIXXICDFXXAKXIFQN----------
1YMT	----------RPTILYALLSPSPR----------
1YUC	----------ASRPAILYALLSSS----------
2Q3Y	----------PAILYALLSS----------
1ZH7	----------SHPTILYTLLS----------

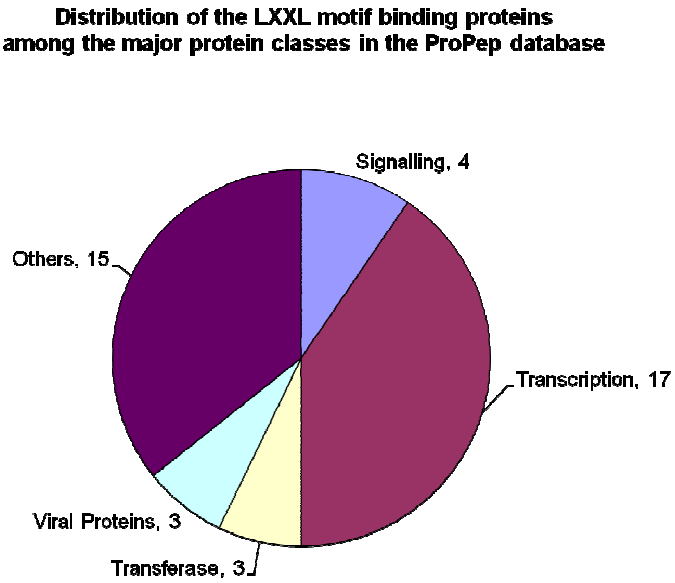

Figure 6-9: The distribution of the LxxL α-helical motif in the ProPep database according to protein classification.

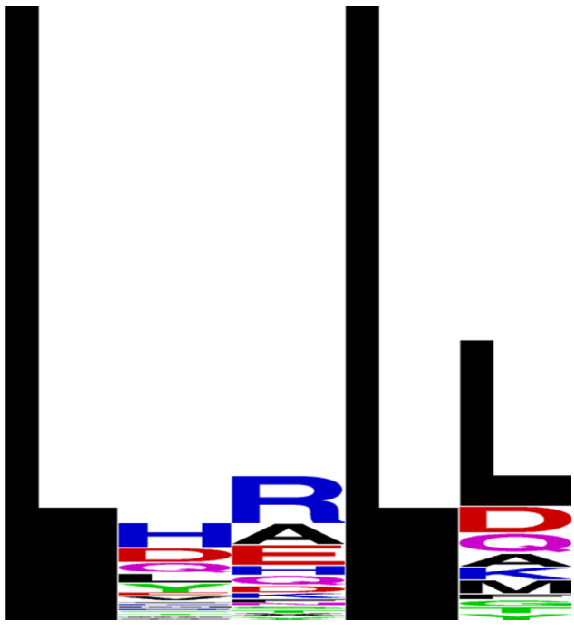

Figure 6-10: The distribution of amino acids at the second, third, and fifth positions of the LxxLx motif shows a high preference for His at position 2, Arg at position 3, and Leu at position 5.

We study the distribution of amino acids in positions 2, 3 (the XX between the two Leus), and 5 (the amino acid after the second Leu) of all the 42 structures in ProPep (Table 6-4). Position 2 is occupied mostly by the polar amino acids Asp, Glu, His, Lys, Met, Gln, Arg, Ser, Thr, Trp, and Tyr. Only seven structures (17%) included one of the hydrophobic amino acids Ala, Leu, and Val at position 2. A similar distribution of amino acids is found at position 3 with only six structures (14%) exhibiting a hydrophobic amino acid. Position 5 exhibits a different distribution, with Leu being a dominant occupier of the fifth slot in 25 structures (60%). Two structures exhibit an Ala residue in the fifth position while the remaining 15 structures include a polar amino acid side chain. Looking closely at positions 2 and 3 (Figure 6-10), the preferred amino acids are His and Arg (26% and 33% respectively). Consequently, the highest occurring LxxL fragment is LHRL (10 occurrences).

6.6 Screening for LHRLL mimics with UFSRAT and Autodock

The LxxLL motif occupies 60% (25/42) of the LxxLx occurrences in ProPep. This motif plays a role as an activator of nuclear receptor functions. It binds and activates the Hepatocyte nuclear factor 4 which regulates the development and physiology of vital organs like the liver, pancreas, and kidney [304]. Mediatory-1 LxxLL motifs play important roles in Estrogen Receptor α-mediated functions in early mammary gland development [305, 306] and the inhibition of this interaction perturbs the localization of the Estrogen Receptor α. Involved in PPARγ and vitamin D receptor interaction [307-310] and mimetics of the LxxLL motif inhibit the interaction of vitamin D receptor with coactivators [311]. As shown in Table 6-4, 10 out of the 25 LxxLL motifs in ProPep are LHRLL (24% of the LxxL motif structures and 2% of

the ProPep database). This section studies these occurrences and documents the virtual screening approach used to find mimetics of the LHRLL motifs.

Table 6-5: The entries in Propep which contain the extended HKILHRLLQ α helical peptide motif. All ten entries are crystal structures of nuclear receptors. The extended motif is marked in green, and the LHRLL motif is marked in yellow.

PDB ID	Function	Peptide Sequence
1NQ7	Retinoid X Receptor: Regulation of gene expression	----------HKILHRLLQE----
1NRL	Pregnane X Receptor (PXR): detection of foreign compounds and regulation of the expression of genes central to drug metabolism and excretion.	CPSSHSSLTERHKILHRLLQEGSPS
1PZL	Gene regulation of nutrient transport and metabolism functions	CPSSHSSLTERHKILHRLLQEGSPS
1UHL	Liver X Receptor: regulation of genes involved lipid homeostasis and inflammation	----------HKILHRLLQD----
1XIU	Retinoid X Receptor: Regulation of gene expression	----------RHKILHRLLQEGSPS
2HC4	Vitamin D Receptor: calcium and phosphorous homeostasis	----------RHKILHRLLQEGSPS
2P1T	Retinoid X Receptor: Regulation of gene expression	----------KHKILHRLLQDSS--
2P54	Peroxisome proliferator activated receptor α: lipid homeostasis	---------ARHKILHRLLQE----
2QZO	Estrogen Receptor α: in bone structural maintenance	----------KHKILHRLLQDSS--
3BEJ	Farnesoid X Receptor: regulation of bile acid and cholesterol homeostasis	CPSSHSSLTERHKILHRLLQEGSPS

The LHRLL motif occurs ten times in the ProPep database. Seven out of these ten occurrences are classified as Transcription proteins. All ten structures are nuclear receptors (Table 6-5). A closer look at those 10 structures reveals that they all include the 9-amino acid motif HKILHRLLQ (Table 6-4 and Table 6-5), significantly larger than the LHRLL motif. This motif was extracted from all ten

structures and superposed, giving a maximum backbone RMSD of 0.3 Å (Figure 6-11). It is hence a structurally conserved motif that binds to nuclear receptors.

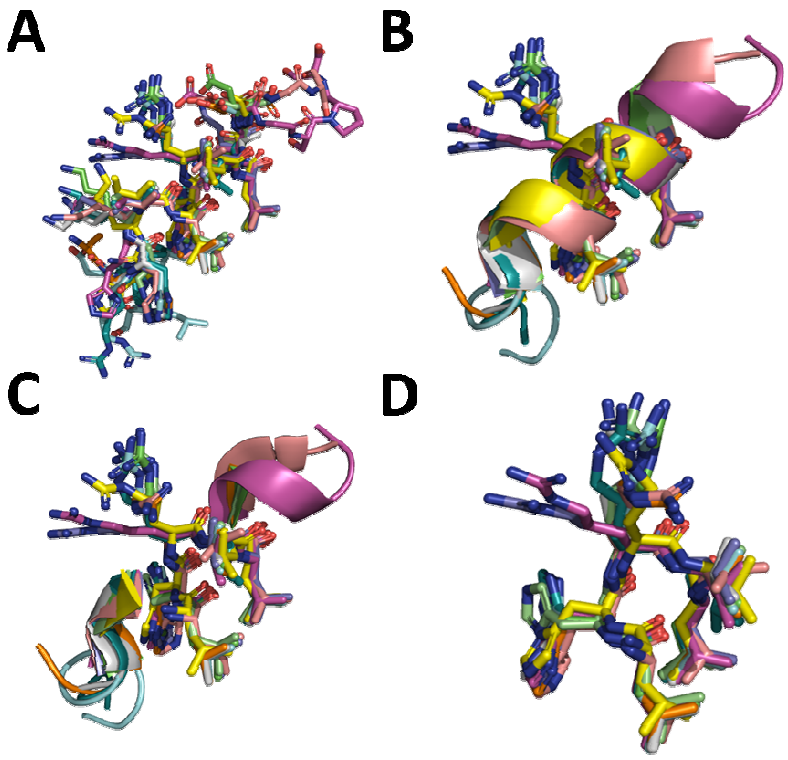

Figure 6-11: The superposition of the HKILHRLLQ motif across the ten nuclear receptors in ProPep (Table 6-5) shows a maximum RMSD of 0.3 Å. (A) shows the full motif in sticks, (B) and (C) show the full motif in cartoons and the LHRLL parts in sticks, and D shows the superposition of the LHRLL segments of this motif. First analysis shows that the Leus in the LHRLL segment are structurally conserved while the Arg and the His side chains undergo slight conformational changes.

We conduct a virtual screening experiment to search for molecules in the EDULISS database that would structurally mimic the behavior of the LHRLL motif. The LHRLL fragment of the PDB structure 1NQ7 is used as a template. The side chains of the His and Arg amino acids are stripped away since they are the least structurally consistent among the superposed 10 nuclear receptor ligands. We use UFSRAT [98] to perform a ligand-based virtual screening experiment and generate the top 50

ligands that would best mimic the structure of the LHRLL motif. The MISCC (Maybridge, InterBioScreen, Specs, Chemdiv, and Chembridge) multiconformer subset of EDULISS is used as a compound repository. This set of oprea filtered [92] compounds contains 3,803,396 multi conformer compounds (conformers generated based on flexibility and number of rotatable bonds in a compound, on average, 4 structures per compound). The top 50 UFSRAT hits are then docked into the pockets of the 10 nuclear receptor structures and the results are analyzed.

Table 6-6: The top 50 generated compounds by UFSRAT as mimetics to the LHRLL α helical motif.

EDULISS ID	UFSRAT Similarity Score	EDULISS ID	UFSRAT Similarity Score
29SPH1-353-024_4	0.755	8SPH1-000-855_3	0.729
8SPH1-142-271_7	0.746	23SPH1-032-299_3	0.729
9SPH1-064-934_7	0.745	29SPH1-314-502_3	0.729
9SPH1-405-156_2	0.744	29SPH1-375-461_3	0.728
8SPH1-266-700_3	0.741	29SPH1-375-461_2	0.728
29SPH1-380-602_4	0.740	9SPH1-402-038_3	0.728
8SPH1-287-949_4	0.740	8SPH1-166-815_3	0.728
8SPH1-142-775_2	0.737	29SPH1-207-667_2	0.728
8SPH1-064-672_2	0.737	8SPH1-197-458_4	0.728
8SPH1-381-204_2	0.737	9SPH1-373-573_5	0.726
8SPH1-275-957_3	0.736	9SPH1-086-496_2	0.726
8SPH1-346-796_1	0.735	29SPH1-375-355_7	0.725
29SPH1-171-482_3	0.734	9SPH1-373-573_6	0.725
29SPH1-233-702_1	0.734	29SPH1-375-347_8	0.725
9SPH1-399-879_5	0.734	29SPH1-129-202_3	0.725
8SPH1-280-623_4	0.733	8SPH1-354-157_3	0.725
8SPH1-283-972_4	0.733	8SPH1-259-804_2	0.725
23SPH1-211-789_2	0.732	29SPH1-380-503_4	0.724
8SPH1-242-443_4	0.732	29SPH1-152-819_2	0.724
9SPH1-110-586_4	0.731	8SPH1-170-401_2	0.724
29SPH1-375-461_5	0.731	8SPH1-287-949_5	0.724
29SPH1-375-432_5	0.731	9SPH1-354-347_2	0.723
29SPH1-375-355_5	0.730	9SPH1-189-664_2	0.723
9SPH1-385-831_2	0.729	29SPH1-375-355_4	0.723
29SPH1-375-445_3	0.729	29SPH1-375-313_2	0.723

The top 50 UFSRAT hits were docked into the LHRLL binding domains of the ten nuclear receptor structures in the ProPep database using an adapted version of the docking program AutoDock [71]. This version has been adapted to handle high-throughput virtual screening experiments, by docking a group of molecules to the same receptor using parallel computing (adaptation done by Dr Douglas Houston). This version of AutoDock then analyzes the docking results for all the screened compounds and ranks these candidates by the predicted binding energy to the target. The performance of these compounds against the ten nuclear receptors was analyzed (Table 6-7 and Table 6-8). The best results were observed with the structures 1NRL (Pregnane X Receptor), 2HC4 (Vitamin D Receptor), 3BEJ (Farnesoid X Receptor), 2P54 (Peroxisome Receptor α), and 1UHL (Liver X Receptor). That being said, the performance of these 50 compounds against the other nuclear receptors is also good; producing estimated affinities in the low micromolar region (Table 6-7 and Table 6-8).

Table 6-7: The estimated binding affinity (by AutoDock) of the top 50 LHRLL peptidomimetic compounds (generated by UFSRAT) against the ten LHRLL binding nuclear receptors in ProPep. All sub-μM affinities are marked in bold font (Part 1/2).

Compound	1NQ7	1NRL	1PZL	1UHL	1XIU
8-170-401_2	4.33μM	1.22μM	8.70μM	2.89μM	1.85μM
8-197-458_4	20.69μM	**772.69pM**	1.65mM	**74.17nM**	62.99μM
9-086-496_2	12.28μM	**950.51pM**	415.60μM	**124.88nM**	45.68μM
8-142-775_2	8.14μM	**637.92pM**	551.37μM	**131.27nM**	32.56μM
8-381-204_2	7.87μM	24.27μM	32.44μM	23.77μM	13.87μM
8-242-443_4	5.93μM	4.00μM	20.10μM	17.46μM	6.19μM
29-353-024_4	61.93μM	**831.42nM**	66.62μM	64.74μM	16.86μM
29-375-347_8	16.23μM	13.22μM	49.93μM	7.62μM	11.70μM
29-380-602_4	6.16μM	16.92μM	60.82μM	12.11μM	12.16μM
23-032-299_3	162.87μM	**2.38nM**	2.20mM	**418.37nM**	447.34μM
9-354-347_2	62.69μM	5.70μM	63.63μM	26.48μM	11.43μM
8-000-855_3	215.66μM	**2.47nM**	6.01mM	**447.09nM**	390.05μM
29-375-445_3	81.68μM	12.02μM	223.08μM	20.61μM	12.92μM
29-380-503_4	12.47μM	22.12μM	82.94μM	23.56μM	13.25μM
29-152-819_2	107.16μM	**2.85nM**	6.51mM	1.58μM	261.17μM

9-402-038_3	77.13µM	28.78µM	104.77µM	17.47µM	12.73µM
29-314-502_3	26.14µM	13.37µM	75.96µM	23.54µM	16.47µM
8-346-796_1	18.50µM	18.85µM	396.98µM	35.17µM	56.74µM
9-189-664_2	229.99µM	**13.30nM**	3.26mM	1.41µM	1.61mM
29-207-667_2	156.97µM	22.68µM	129.55µM	37.97µM	11.54µM
29-375-355_5	56.48µM	41.58µM	175.87µM	44.67µM	59.55µM
29-375-432_5	52.58µM	81.38µM	185.98µM	70.74µM	52.58µM
8-259-804_2	423.76µM	**9.80nM**	2.79mM	1.22µM	411.91µM
9-064-934_7	53.64µM	55.28µM	137.19µM	84.53µM	14.08µM
29-375-355_4	62.36µM	143.60µM	316.34µM	60.80µM	45.73µM
29-171-482_3	62.04µM	51.38µM	188.98µM	16.51µM	41.19µM
29-375-461_3	102.04µM	81.45µM	206.19µM	51.67µM	63.96µM
9-373-573_6	42.30µM	23.51µM	156.41µM	99.99µM	33.59µM
9-110-586_4	9.19µM	69.97µM	176.43µM	86.69µM	49.11µM
29-375-355_7	86.61µM	71.76µM	191.67µM	50.91µM	62.43µM
29-375-461_2	99.45µM	65.52µM	213.93µM	98.42µM	56.02µM
9-373-573_5	34.00µM	34.62µM	345.90µM	85.11µM	24.71µM
29-375-461_5	79.49µM	45.80µM	303.09µM	97.56µM	65.29µM
8-166-815_3	144.75µM	4.23µM	349.68µM	170.30µM	65.39µM
29-375-313_2	103.94µM	74.13µM	343.55µM	61.23µM	66.04µM
8-142-271_7	69.54µM	47.01µM	143.52µM	91.33µM	23.44µM
8-275-957_3	72.09µM	50.85µM	158.58µM	56.52µM	95.74µM
8-280-623_4	62.07µM	99.85µM	292.78µM	61.53µM	48.97µM
8-283-972_4	62.82µM	17.99µM	161.85µM	116.14µM	61.13µM
29-233-702_1	207.01µM	73.91µM	593.75µM	17.70µM	154.52µM
8-064-672_2	212.34µM	60.42µM	410.26µM	367.71µM	56.80µM
9-399-879_5	216.68µM	123.88µM	484.84µM	123.19µM	106.54µM
8-266-700_3	105.81µM	22.13µM	756.89µM	132.89µM	34.52µM
8-354-157_3	147.56µM	124.68µM	318.08µM	288.59µM	117.20µM
8-287-949_4	162.81µM	80.46µM	691.72µM	156.08µM	65.67µM
8-287-949_5	111.34µM	67.14µM	918.84µM	94.04µM	288.89µM
23-211-789_2	404.41µM	62.82µM	903.24µM	389.97µM	296.13µM
9-385-831_2	642.08µM	134.88µM	586.42µM	383.62µM	79.49µM
29-129-202_3	421.63µM	387.26µM	1.90mM	727.76µM	138.17µM
9-405-156_2	1.06 mM	565.57µM	1.95mM	921.25µM	334.00µM
Lowest Ki	1.06mM	565.57µM	6.51mM	921.25µM	1.61mM
Highest Ki	4.33µM	772.69pM	8.7µM	74.17µM	1.85µM

Table 6-8: The estimated binding affinity (by AutoDock) of the top 50 LHRLL peptidomimetic compounds (generated by UFSRAT) against the ten LHRLL binding nuclear receptors in ProPep. All sub-μM affinities are marked in bold font (Part 2/2).

Compound	2HC4	2P1T	2P54	2QZO	3BEJ
8-170-401_2	**987.94nM**	1.14μM	1.51μM	1.60μM	1.14μM
8-197-458_4	**15.24nM**	61.07μM	37.40μM	3.74μM	8.48μM
9-086-496_2	**19.18nM**	112.16μM	50.05μM	8.03μM	8.27μM
8-142-775_2	**22.82nM**	94.72μM	75.55μM	9.98μM	10.07μM
8-381-204_2	2.06μM	7.91μM	1.07μM	7.58μM	3.23μM
8-242-443_4	13.41μM	9.60μM	4.10μM	17.27μM	5.61μM
29-353-024_4	**826.78nM**	2.90μM	9.05μM	31.48μM	5.10μM
29-375-347_8	4.32μM	2.47μM	15.78μM	31.23μM	4.58μM
29-380-602_4	9.35μM	8.13μM	6.75μM	54.57μM	3.60μM
23-032-299_3	**76.95nM**	395.45μM	89.79μM	17.00μM	62.74μM
9-354-347_2	18.07μM	10.99μM	7.24μM	74.92μM	6.52μM
8-000-855_3	**112.51nM**	399.27μM	91.16μM	30.30μM	82.11μM
29-375-445_3	4.33μM	8.03μM	20.32μM	66.10μM	4.11μM
29-380-503_4	15.92μM	21.14μM	9.03μM	51.56μM	12.58μM
29-152-819_2	**52.54nM**	539.65μM	141.66μM	45.09μM	129.48μM
9-402-038_3	18.81μM	23.94μM	10.48μM	22.57μM	13.96μM
29-314-502_3	29.48μM	27.59μM	13.16μM	60.40μM	15.03μM
8-346-796_1	28.01μM	43.65μM	5.75μM	45.84μM	8.28μM
9-189-664_2	**120.19nM**	1.76mM	182.57μM	123.32μM	59.29μM
29-207-667_2	46.61μM	17.18μM	13.70μM	99.23μM	39.94μM
29-375-355_5	7.25μM	48.77μM	28.97μM	185.20μM	4.52μM
29-375-432_5	6.26μM	53.46μM	13.93μM	88.91μM	8.17μM
8-259-804_2	**204.56nM**	1.50mM	404.30μM	97.61μM	141.42μM
9-064-934_7	14.03μM	21.70μM	9.67μM	189.72μM	77.37μM
29-375-355_4	8.21μM	28.68μM	14.65μM	224.45μM	5.42μM
29-171-482_3	52.26μM	43.40μM	22.22μM	65.11μM	36.15μM
29-375-461_3	12.69μM	29.92μM	19.40μM	146.10μM	10.24μM
9-373-573_6	93.58μM	38.84μM	28.29μM	93.91μM	17.17μM
9-110-586_4	42.27μM	77.43μM	32.61μM	94.48μM	19.38μM
29-375-355_7	8.79μM	49.45μM	55.02μM	199.17μM	7.03μM
29-375-461_2	9.50μM	55.54μM	25.33μM	157.81μM	12.46μM
9-373-573_5	63.47μM	41.23μM	27.24μM	77.11μM	52.93μM
29-375-461_5	11.16μM	62.78μM	47.47μM	165.06μM	7.98μM
8-166-815_3	19.10μM	257.28μM	45.27μM	96.98μM	8.31μM
29-375-313_2	13.61μM	97.84μM	22.61μM	434.27μM	12.85μM
8-142-271_7	97.13μM	37.87μM	44.97μM	215.15μM	58.87μM
8-275-957_3	15.37μM	90.74μM	66.78μM	258.01μM	52.68μM
8-280-623_4	21.58μM	82.19μM	57.56μM	250.04μM	49.76μM
8-283-972_4	23.69μM	196.39μM	51.91μM	390.38μM	74.82μM
29-233-702_1	41.61μM	113.37μM	14.14μM	152.72μM	37.40μM
8-064-672_2	48.03μM	64.54μM	23.31μM	100.83μM	54.19μM
9-399-879_5	78.94μM	22.05μM	9.53μM	312.22μM	106.27μM
8-266-700_3	143.02μM	117.61μM	54.14μM	286.66μM	98.80μM
8-354-157_3	60.88μM	96.99μM	23.14μM	140.68μM	62.25μM

8-287-949_4	120.04µM	58.61µM	12.66µM	767.42µM	136.85µM
8-287-949_5	90.33µM	57.91µM	57.04µM	396.47µM	66.71µM
23-211-789_2	75.86µM	182.77µM	40.14µM	862.14µM	236.44µM
9-385-831_2	47.80µM	131.48µM	228.39µM	600.86µM	240.50µM
29-129-202_3	181.81µM	170.49µM	270.11µM	119.85µM	398.11µM
9-405-156_2	421.34µM	262.29µM	582.18µM	1.08mM	906.11µM
Lowest Ki	421.34µM	1.76mM	582.18µM	1.08mM	906.11µM
Highest Ki	15.24nM	1.14µM	1.07 µM	1.60µM	1.14µM

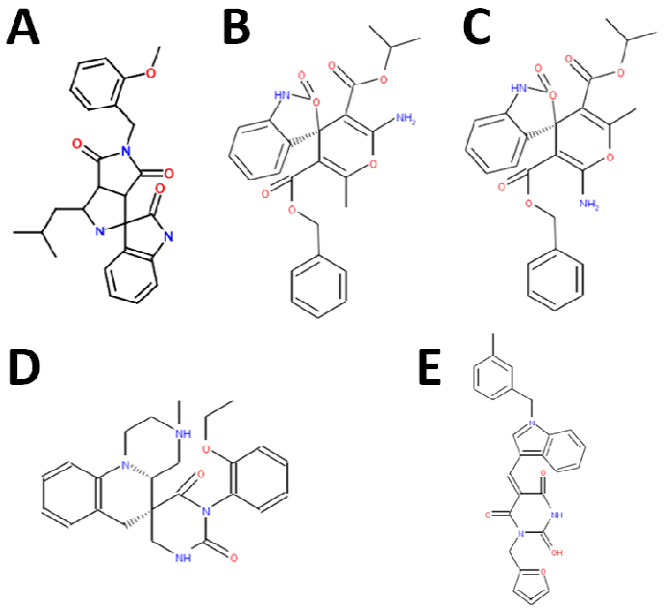

Figure 6-12: The five top scoring compounds mimicking the interaction between the LHRLL α helical motif and the ten nuclear receptor structures in the ProPep database. Compounds are scored based on the average affinity of interaction with the ten nuclear receptors.

Table 6-9: The binding energies of the best 5 LHRLL peptidomimetic compounds (Figure 6-12) against the ten nuclear receptors (Part 1/2).

Compound	1NQ7	1NRL	1PZL	1UHL	1XIU
A	4.33 µM	1.22 µM	8.70 µM	2.89 µM	1.85 µM
B	20.69 µM	772.69 pM	1.65 mM	74.17 nM	62.99 µM
C	12.28 µM	950.51 pM	415.60 µM	124.88 nM	45.68 µM
D	7.87 µM	24.27 µM	32.44 µM	23.77 µM	13.87 µM
E	5.93 µM	4.00 µM	20.10 µM	17.46 µM	6.19 µM

Table 6-10: The binding energies of the best 5 LHRLL peptidomimetic compounds (Figure 6-12) against the ten nuclear receptors (Part 2/2).

Compound	2HC4	2P1T	2P54	2QZO	3BEJ
A	987.94 nM	1.14 µM	1.51 µM	1.60 µM	1.14 µM
B	15.24 nM	61.07 µM	37.40 µM	3.74 µM	8.48 µM
C	19.18 nM	112.16 µM	50.05 µM	8.03 µM	8.27 µM
D	2.06 µM	7.91 µM	1.07 µM	7.58 µM	3.23 µM
E	13.41 µM	9.6 µM	4.1 µM	17.72 µM	5.61 µM

Out of the 50 generated compounds, 5 stand out, having the highest average binding affinity to the LHRLL binding sites of the ten nuclear receptors in the ProPep database (Figure 6-12, Table 6-9, and Table 6-10). These compounds are referred to as compounds A through E, as per Figure 6-12. The predicted binding affinities of these compounds towards each of the nuclear receptors ranged between 773 picoMolar and 416 microMolar. However, most of the predicted affinities lie in the high nanoMolar to low microMolar region. The binding of compound B to the Pregnane X Receptor (1NRL) at a predicted affinity of 773 pM is solidified by a network of 4 hydrogen bonds in addition to the hydrophobic stacking between the molecule and the protein (Figure 6-13). Similar interaction profiles are noticed for the other UFSRAT hits as well (Figure 6-14).

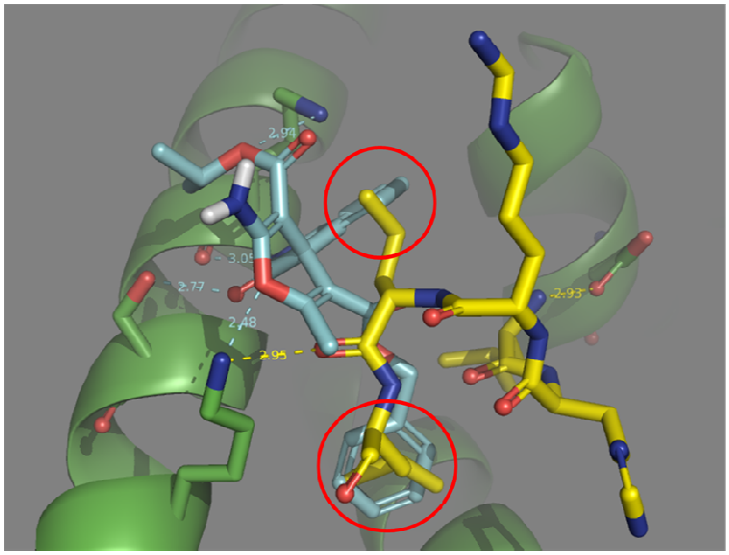

Figure 6-13: The binding of compound B (cyan) to the Pregnane X Receptor (1NRL) at a predicted affinity of 772.69 pM. The LHRLL motif is shown in yellow, making 2 hydrogen bonds (yellow) with 1NRL. The benzene rings of compound B fill in the area occupied by the Leu side chains of the LHRLL motif (red circles) and its interaction is solidified by 4 hydrogen bonds (in cyan).

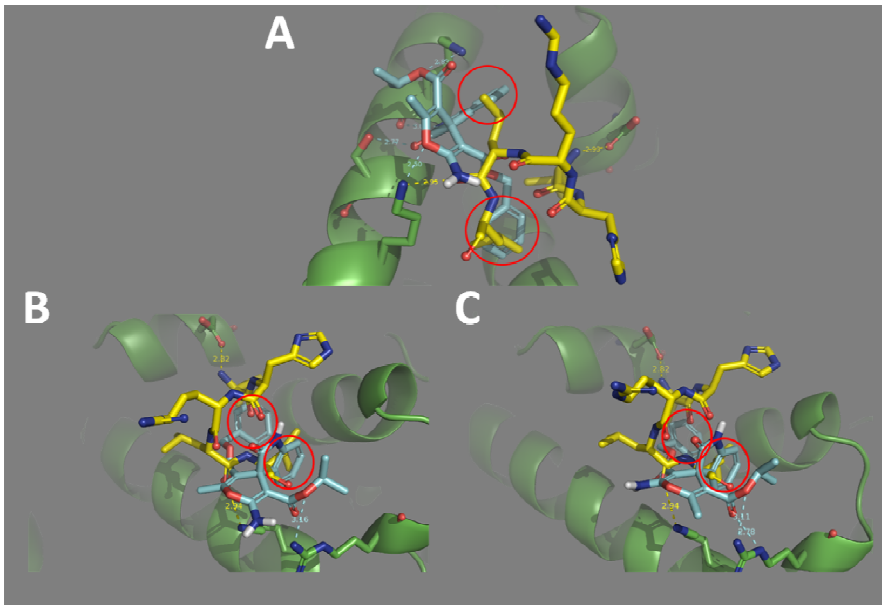

Figure 6-14: The interaction with top scoring UFSRAT hits (cyan sticks) with nuclear receptors (cartoons). The original LHRLL motifs are shown for comparison (yellow sticks). (A) shows the binding of compound C to the pregnane X receptor (pdb 1NRL) at a predicted affinity of 950 pM, (B) shows the binding of compound B to the vitamin D receptor (pdb 2HC4) at a predicted affinity of 15 nM, and (C) shows the binding of compound C to the vitamin D receptor (pdb 2HC4) at a predicted affinity of 19 nM. The benzene rings of the UFSRAT hits fill in the areas occupied by the Leu side chains of the LHRLL motif (red circles). The interactions are further solidified with hydrogen bonds.

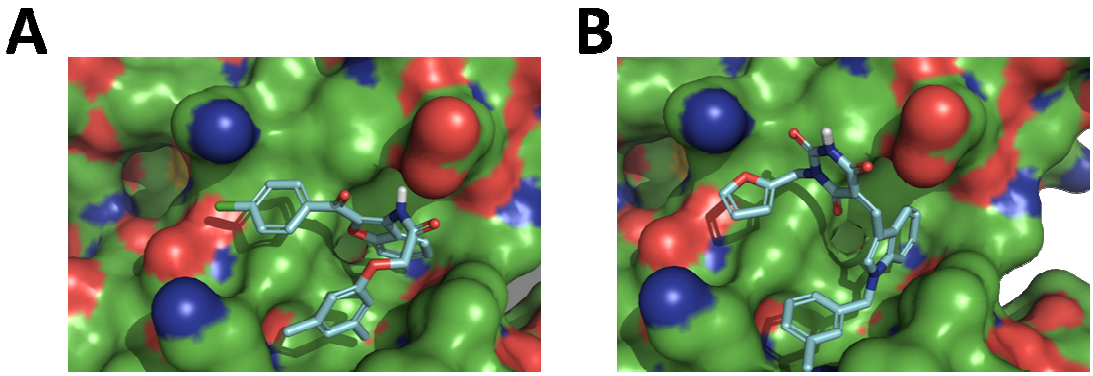

Figure 6-15: The binding sites of the nuclear receptors (the pregnane X receptor, pdb 1NRL is shown) are mainly hydrophobic pockets. This justifies the favouring of benzene rings and their halogenated and methylated derivatives (A). The existence of a few polar atoms resulted in the substitution of some ring carbons with nitrogens or oxygens to facilitate hydrogen bonding (B).

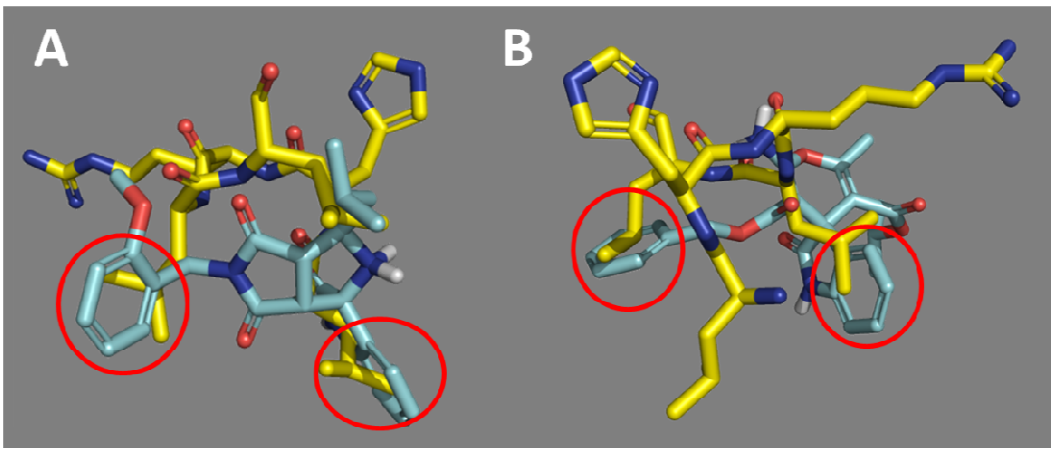

Figure 6-16: The AutoDock simulations showed that benzene rings are good substitutions for the hydrophobic Leu side chains. Figures (A) and (B) show the overlap of the docked compounds B and C (Figure 6-12) with the LHRLL motif. Structures shown are docked to the Pregnane X Receptor, pdb 1NRL.

The ligand binding sites of the nuclear receptors are mainly hydrophobic pockets. A few polar atoms exist but that is not enough to change the hydrophobicity of the binding sites (Figure 6-15). This is in agreement with the binding of the hydrophobic Leu side chains in the pockets. Indeed, the UFSRAT/AutoDock screening for

LHRLL mimetics generated compounds with mainly hydrophobic substructures. The hydrophobic Leu side chains are often substituted with ring substructures: 44 out of the 50 docked molecules to 1NRL resulted in a benzene ring or one of its derivatives (eg naphthyl) occupying the place of the Leu side chain (Figure 6-16). Benzene rings and their derivatives have been selected as good mimetics for the Leu side chains via the UFSRAT shape and atom type matching algorithm. AutoDock simulations backed up this selection by revealing strong interactions between these ring substructures and the proteins. In some cases, the carbons on these ring substructures are substituted by nitrogens, oxygens, carbonyl, or halogenated carbons. These substitutions allowed for the creation of hydrogen bonds between the compounds and the protein, enhancing the binding energy.

Although the predictions produced in this chapter are virtual results and have to be validated with wet-lab experiments, the results of this experiment are still important. UFSRAT [98] and Autodock [71] are well established programs known for their success in virtual screening. Careful examination of the results showed that the interaction between the generated hits and the nuclear receptors is solid (Figure 6-13). The selected hits react similarly with the nuclear receptors and this is assuring as all ten structures bind the same LHRLL motif. The actual binding affinities will not be as high as predicted by the virtual programs (AutoDock predicted picoMolar affinity for the interaction between compound B the Pregnane X Receptor), but we are still confident of a strong binding between the two molecules. By using all these carefully crafted programs and combining them in a synergetic and self validating method, we have managed to reduce a space of testable compounds from millions to

tens, or even hundreds. With the parallelization of most of our systems, the time spent on running the virtual screening experiments is less than two weeks. In just that time, computational drug discovery is capable of providing insights on inhibiting the LHRLL motif : nuclear receptor interaction.

7 Conclusion

7.1 Summary

This work contributes to protein based, ligand based, and database based drug discovery. The STP method (Chapter 2) plays an important role in detecting protein binding sites. This functionality serves an important role in minimizing the search space for docking algorithms by creating pseudo ligands used as markers for the locations of binding sites. Chapter 3 demonstrates the power of this method in predicting allosteric binding sites, enzyme commission classes, and ranking docking orientations. Chapter 4 discusses the chemical and topological characteristics of binding sites and compares the binding sites that bind ligands, peptides, and protein domains. This study shows the importance of polar interactions in protein-ligand binding only while hydrophobic interactions play important roles in all types of interaction. The statistical binding preference between STP triplets and ligand atom types has also been studied, quantified, and converted to statistical free energy of binding between ligand atoms and surface triplets.

Chapter 5 discusses a different aspect of computational drug discovery, and focuses on ligand-based drug discovery. It outlines the use of a ligand (Baff-R) with computational algorithms to design a virtual library of molecules that are likely to mimic the interactions of this Baff-R ligand. It also outlines the use of docking and structural similarity algorithms (Autodock and UFSRAT) to study protein interaction and hence refine the molecules included in the virtual library. Finally, Chapter 6 discusses the use of interaction databases to detect interaction motifs and then the

virtual screening for molecules that are likely to mimic the interactions performed by this motif. Consequently the ProPep database has been used to identify the LHRLL alpha helical motif that binds to nuclear receptors. Docking and structural similarity methods (Autodock and UFSRAT) are then used to create a virtual library of possible inhibitors for the interaction of the LHRLL motif with nuclear receptors. Virtual experiments show that the molecules bind consistently to all the ten receptors.

7.2 Future Work

The results and methods introduced in this work pave the way towards further advancements including:

- Creating an STP profile for allosteric binding sites
- Creating an STP profile for metal binding sites
- Using the STP coloring routine on snapshots of molecular dynamics simulations to show the formation and concealing of binding hotspots
- Creating STP profiles with varied probe molecule size (this work used 1.4 Å). This should allow for assessment of the protein surface at different resolutions. Using smaller probes would account for atoms located at the bottom of pockets (such atoms are likely to be missed by larger probes). Using larger probes would define the protein surface from the point of view of non water molecules (e.g. carbon).
- Using the statistical free energy of interaction between ligand atoms and surface triplets to score and rank small molecule docking results
- Using the statistical free energy of interaction between ligand atoms and surface triplets to create maps showing the best choices of atom types at certain locations around the protein. These maps can then be used manually or computationally to construct ligands that would bind to the protein with high affinity

- Using the triplet composition of a binding site to identify its function (expansion on the EC number prediction to cover a wide range of functions)
- In-vitro testing and optimization of the high scoring compounds genertated by the virtual screening simulations for CRK3 (Chapter 3)
- In-vitro testing and optimization of the high scoring cyclic peptides genertated by the virtual screening simulations for BLYS (Chapter 5)
- In-vitro testing and optimization of the high scoring compounds genertated by the virtual screening simulations for the Nuclear receptors (Chapter 6).

7.3 The Age of Multiplicity

Despite the current advances, our understanding of molecular interaction is still limited. Different models try to simulate real interactions in various methods and sometimes reach contradicting results. The increasing power of computing is still insufficient to perform detailed calculations quickly (molecular dynamics simulations require days and sometimes weeks to finish the simulation). As a result, all computer simulations are mere approximations of reality and rely on heuristics extensively to perform their predictions. Nevertheless, the input provided by these simulations is very influential. Using the results of many simulations by multiple programs and trying to reach a consensus is currently the best approach to generate reliable results. Rentzsch and Orengo (2009) [53] define this as the *age of multiplicity*, where the use of multiple tools and multiple starting points to study a certain phenotype is necessary to generate a more accurate picture of the general mechanism.

8 References

1. Padlan, E.A., et al., *Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex.* Proc Natl Acad Sci USA, 1989. **86**(15): p. 5938-42.
2. Su, H.P., et al., *Structural basis for the inhibition of RNase H activity of HIV-1 reverse transcriptase by RNase H active site-directed inhibitors.* J Virol, 2010. **84**(15): p. 7625-33.
3. Biertümpfel, C., et al., *Structure and mechanism of human DNA polymerase eta.* Nature, 2010. **465**(7301): p. 1044-8.
4. Pérot, S., et al., *Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery.* Drug Discov Today, 2010.
5. Vaz, R.J., et al., *The challenges of in silico contributions to drug metabolism in lead optimization.* Expert Opin Drug Metab Toxicol, 2010. **6**(7): p. 851-61.
6. Colizzi, F., et al., *Single-molecule pulling simulations can discern active from inactive enzyme inhibitors.* J Am Chem Soc, 2010. **132**(21): p. 7361-71.
7. Giancarlo, R., D. Scaturro, and F. Utro, *Textual data compression in computational biology: a synopsis.* Bioinformatics, 2009. **25**(13): p. 1575-86.
8. Peng, H., *Bioimage informatics: a new area of engineering biology.* Bioinformatics, 2008. **24**(17): p. 1827-36.
9. Zhang, Y., M.E. Devries, and J. Skolnick, *Structure modeling of all identified G protein-coupled receptors in the human genome.* PLoS Comput Biol, 2006. **2**(2): p. 13.
10. Guvench, O. and A.D.J. MacKerell, *Computational fragment-based binding site identification by ligand competitive saturation.* PLoS Comput Biol, 2009. **5**(7): p. e1000435.
11. Nagamine, N., et al., *Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening.* PLoS Comput Biol, 2009. **5**(6): p. e1000397.
12. Suderman, M. and M. Hallett, *Tools for visually exploring biological networks.* Bioinformatics, 2007. **23**(20): p. 2651-9.
13. Bock, C. and T. Lengauer, *Computational epigenetics.* Bioinformatics, 2008. **24**(1): p. 1-10.
14. Moore, G.E., *Lithography and the Future of Moore's Law.* Optical/Laser Microlithography VIII: Proceedings of the SPIE, 1995. **2440**: p. 2-17.
15. Manavski, S.A. and G. Valle, *CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment.* BMC Bioinformatics, 2008. **9**(Suppl 2): p. S10.

16. Macindoe, G., et al., *HexServer: an FFT-based protein docking server powered by graphics processors.* Nucleic Acids Res, 2010. **38**(Web Server Issue): p. W445-9.

17. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

18. Lo Conte, L., et al., *SCOP database in 2002: refinements accommodate structural genomics.* Nucleic Acids Res, 2002. **30**(1): p. 264-7.

19. Orengo, C.A., et al., *The CATH protein family database: a resource for structural and functional annotation of genomes.* Proteomics, 2002. **2**(1): p. 11-21.

20. Greene, L.H., et al., *The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution.* Nucleic Acids Res, 2007. **35**(Database Issue): p. D291-7.

21. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.

22. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

23. Pagni, M., et al., *MyHits: improvements to an interactive resource for analyzing protein sequences.* Nucleic Acids Res, 2007. **35**(Web Server Issue): p. W433-7.

24. Notredame, C., *Computing multiple sequence/structure alignments with the T-coffee package.* Curr Protoc Bioinformatics, 2010. **3**(8): p. 1-25.

25. Higgins, D.G., J.D. Thompson, and T.J. Gibson, *Using CLUSTAL for multiple sequence alignments.* Methods Enzymol, 1996. **266**: p. 383-402.

26. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2010. **38**(Database issue): p. D5-16.

27. Sander, C. and R. Schneider, *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins, 1991. **9**(1): p. 56-68.

28. Rodriguez, R. and G. Vriend, *Professional gambling*. Procedings of the NATO Advanced Study Institute on Biomolecular Structure and Dynamics: Recent Experimental and Theoretical Advances., 1997.

29. Magloire, A., *Grep: Searching for a Pattern*. 2000, Bloomington, IN: iUniverse Inc.

30. Chomsky, N., *Syntactic Structures*. 2 ed. 2002, Berlin: Mouton.

31. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad of Sci U S A, 1992. **89**(22): p. 10915-9.

32. Dayhoff, M.O., R. Schwartz, and B.C. Orcutt, *A model of Evolutionary Change in Proteins*. Atlas of protein sequence and structure. Vol. 5. 1978: National Biomedical Research Foundation.

33. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J, 1986. **5**(4): p. 823-6.

34. Murzin, A.G., *How far divergent evolution goes in proteins.* Curr Opin Struct Biol, 1998. **8**(3): p. 380-7.

35. Lemmen, C. and T. Lengauer, *Computational methods for the structural alignment of molecules.* J Comput Aided Mol Des, 2000. **14**(3): p. 215-32.

36. Zhu, J. and Z. Weng, *FAST: a novel protein structure alignment algorithm.* Proteins, 2005. **58**(3): p. 618-27.

37. Shatsky, M., R. Nussinov, and H.J. Wolfson, *A method for simultaneous alignment of multiple protein structures.* Proteins, 2004. **56**(1): p. 143-56.

38. Maiti, R., et al., *SuperPose: a simple server for sophisticated structural superposition.* Nucleic Acids Res, 2004. **24**(Web Server issue): p. W590-4.

39. Fischer, D., et al., *Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding.* Protein Sci, 1994. **3**(5): p. 769-78.

40. Wallace, A.C., N. Borkakoti, and J.M. Thornton, *TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites.* Protein Sci, 1997. **6**(11): p. 2308-23.

41. May, A.C. and M.S. Johnson, *Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions.* Protein Eng, 1995. **8**(9): p. 873-82.

42. Diederichs, K., *Structural superposition of proteins with unknown alignment and detection of topological similarity using a six-dimensional search algorithm.* Proteins, 1995. **23**(2): p. 187-95.

43. Delano, W.L., *The PyMOL Molecular Graphics System, Delano Scientific, Palo Alto, CA, USA,* http://www.pymol.org. 2002.

44. Pettersen, E.F., et al., *UCSF Chimera-a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.

45. Accelrys, *Accelrys DS Visualiser*. 2005.

46. Jones, S. and J.M. Thornton, *Searching for functional sites in protein structures.* Curr Opin Chem Biol, 2004. **8**(1): p. 3-7.

47. Nagano, N., C.A. Orengo, and J.M. Thornton, *One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.* J Mol Biol, 2002. **321**(5): p. 741-65.

48. Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson, *Recognition of functional sites in protein structures.* J Mol Biol, 2004. **339**(3): p. 607-33.

49. Duhovny, D., R. Nussinov, and H.J. Wolfson, *Efficient unbound docking of rigid molecules.* Proceedings of the Second Workshop on Algorithms in

Bioinformatics. Lecture Notes in Computer Science (Guigo, R. and Gusfield, D., eds) September 16-21, 2002, Rome, Italy, 2002. **2452**: p. 185-200.

50. Rosen, M., et al., *Molecular shape comparisons in searches for active sites and functional similarity.* Protein Eng, 1998. **11**(4): p. 263-77.

51. Kinoshita, K. and H. Nakamura, *Identification of protein biochemical functions by similarity search using the molecular surface database eF-site.* Protein Sci, 2003. **12**(8): p. 1589-95.

52. Schmitt, S., D. Kuhn, and G. Klebe, *A new method to detect related function among proteins independent of sequence and fold homology.* J Mol Biol, 2002. **323**(2): p. 387-406.

53. Rentzsch, R. and C.A. Orengo, *Protein function prediction--the power of multiplicity.* Trends Biotechnol, 2009. **27**(4): p. 210-9.

54. Martin, D.M., B. M, and G.J. Barton, *GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes.* BMC Bioinformatics, 2004. **5**.

55. Hawkins, T., et al., *PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data.* Proteins, 2009. **74**(3): p. 566-82.

56. Jones, C.E., et al., *GOSLING: a rule-based protein annotator using BLAST and GO.* Bioinformatics, 2008. **24**(22): p. 2628-9.

57. Engelhardt, B.E., et al., *Protein molecular function prediction by Bayesian phylogenomics.* PLoS Comput Biol, 2005. **1**(5): p. e45.

58. Jöcker, A., et al., *Protein function prediction and annotation in an integrated environment powered by web services (AFAWE).* Bioinformatics, 2008. **24**(20): p. 2393-4.

59. Mulder, N.J., et al., *In Silico Characterization of Proteins: UniProt, InterPro and Integr8.* Mol Biotechnol, 2008. **38**(2): p. 165-77.

60. Hulo, N., et al., *The 20 years of PROSITE.* Nucleic Acids Res, 2008. **36**(Database Issue): p. D245-9.

61. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database Issue): p. D211-22.

62. Kaplan, N., et al., *ProtoNet 4.0: a hierarchical classification of one million protein sequences.* Nucleic Acids Res, 2005. **33**(Database issue): p. D216-8.

63. Alexeyenko, A., et al., *Automatic clustering of orthologs and inparalogs shared by multiple proteomes.* Bioinformatics, 2006. **22**(14): p. e9-15.

64. Jensen, L.J., et al., *Prediction of human protein function according to Gene Ontology categories.* Bioinformatics, 2003. **19**(5): p. 635-42.

65. Cai, C.Z., et al., *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence.* Nucleic Acids Res, 2003. **31**(13): p. 3692-7.

66. Lobley, A., et al., *FFPred: an integrated feature-based function prediction server for vertebrate proteomes.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W297-302.

67. Shen, H.B. and K.C. Chou, *EzyPred: a top-down approach for predicting enzyme functional classes and subclasses.* Biochem Biophys Res Commun, 2007. **364**(1): p. 53-9.

68. Taylor, P., et al., *Bioinformatics Tools for Ligand Discovery*, in *Wellcome Trust ISAB Meeting*. 2009, University of Edinburgh: Edinburgh.

69. Taylor, P., et al., *Ligand discovery and virtual screening using the program LIDAEUS.* Br J Pharmacol, 2008. **153**: p. S55-S67.

70. Clark, M., R.D.I. Cramer, and N. van Opdenbosch, *Validation of the general purpose tripos 5.2 force field.* J Comp Chem, 1989. **10**(8): p. 982-1012.

71. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* J Comput Chem, 1998. **19**(14): p. 1639-62.

72. Ritchie, D.W. and G.J.L. Kemp, *Protein docking using spherical polar Fourier correlations.* Proteins, 2000. **39**(2): p. 178-94.

73. Gray, J.J., et al., *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* J Mol Biol, 2003. **331**(1): p. 281-99.

74. Mashiach, E., R. Nussinov, and H.J. Wolfson, *FiberDock: Flexible induced-fit backbone refinement in molecular docking.* Proteins, 2010. **78**(6): p. 1503-19.

75. Metropolis, N. and S. Ulam, *The Monte Carlo Method.* Journal of the American Statistical Association, 1949. **44**(247): p. 335-341.

76. Andrusier, N., R. Nussinov, and H.J. Wolfson, *FireDock: fast interaction refinement in molecular docking.* Proteins, 2007. **69**(1): p. 139-59.

77. Hinsen, K., *Analysis of domain motions by approximate normal mode calculations.* Proteins, 1998. **33**(3): p. 417-429.

78. Tirion, M., *Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis.* Phys Rev Lett, 1996. **77**(9): p. 1905-8.

79. Ma, J., *New advances in normal mode analysis of supermolecular complexes and applications to structural refinement.* Curr Protein Pept Sci, 2004. **5**(2): p. 119-23.

80. Kingsford, C.L., B. Chazelle, and M. Singh, *Solving and analyzing side-chain positioning problems using linear and integer programming.* Bioinformatics, 2005. **21**(7): p. 1028-36.

81. Inbar, Y., et al., *Prediction of Multimolecular Assemblies by Multiple Docking.* J Mol Biol, 2005. **349**(2): p. 435-47.

82. Dror, O., et al., *EMatch: an efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-EM maps of large macromolecular assemblies.* Acta Crystallogr D Biol Crystallogr, 2007. **63**(Pt 1): p. 42-9.

83. Lasker, K., et al., *Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly.* J Mol Biol, 2009. **388**(1): p. 180-94.

84. Fu, G., et al., *Conformational Changes and Substrate Recognition in Pseudomonas aeruginosa D-Arginine Dehydrogenase.* Biochemistry, 2010(In Press).

85. Matsushima, A., et al., *ERRgamma tethers strongly bisphenol A and 4-alpha-cumylphenol in an induced-fit manner.* Biochem Biophys Res Commun, 2008. **373**(3): p. 408-13.

86. Morgan, H.P., et al., *Allosteric mechanism of pyruvate kinase from Leishmania mexicana uses a rock and lock model.* J Biol Chem, 2010. **285**(17): p. 12892-8.

87. Kirkpatrick, P. and C. Ellis, *Chemical Space.* Nature, 2004. **432**(7019): p. 823.

88. Ekins, S., *Systems-ADME/Tox: resources and network approaches.* J Pharmacol Toxicol Methods, 2006. **53**(1): p. 38-66.

89. Irwin, J.J. and B.K. Shoichet, *ZINC--a free database of commercially available compounds for virtual screening.* J Chem Inf Model, 2005. **45**(1): p. 177-82.

90. Hsin, K., *The development and use of databases for ligand-protein interaction studies. PhD Thesis*, in *School of Biological Sciences*. 2010, University of Edinburgh: Edinburgh.

91. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv Drug Deliv Rev, 2001. **46**(1-3): p. 3-26.

92. Hann, M.M. and T.I. Oprea, *Pursuing the leadlikeness concept in pharmaceutical research.* Curr Opin Chem Biol, 2004. **8**(3): p. 255-63.

93. Olah, M.M., C.G. Bologa, and T.I. Oprea, *Strategies for compound selection.* Curr Drug Discov Technol, 2004. **1**(3): p. 211-20.

94. Congreve, M., et al., *A 'rule of three' for fragment-based lead discovery?* Drug Discov Today, 2003. **8**(19): p. 876-7.

95. Eswar, N., et al., *Comparative Protein Structure Modeling With MODELLER.* Curr Protoc Protein Sci, 2007. **2**(9).

96. Mehio, W., et al., *Identification of Protein Binding Surfaces using Surface Triplet Propensities.* Bioinformatics, 2010. **In Press**.

97. Grant, J.A., M.A. Gallardo, and B.T. Pickup, *A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape.* J Comp Chem, 1996. **17**(14): p. 1653-66.

98. Shave, S., *High Performance Structure and Ligand Based Virtual Screening. PhD Thesis*, in *School of Biological Sciences*. 2010, University of Edinburgh Edinburgh.

99. Vanhee, P., et al., *PepX: a structural database of non-redundant protein-peptide complexes.* Nucleic Acids Res, 2010. **38**(Database Issue): p. D545-51.

100. Kundrotas, P.J., Z. Zhu, and I.A. Vakser, *GWIDD: Genome-wide protein docking database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D513-7.

101. Wang, R., et al., *The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.* J Med Chem, 2004. **2004**(47): p. 12.

102. Wang, R., et al., *The PDBbind database: methodologies and updates.* J Med Chem, 2005. **48**(12): p. 4111-9.

103. Harding, S., *Database Analysis of Protein-Peptide Interactions and In Silico Screening for Peptidomimetics. Phd Thesis*, in *School of Biological Sciences*. 2008, University of Edinburgh: Edinburgh.

104. Matyas, G., et al., *Identification and in silico analyses of novel TGFBR1 and TGFBR2 mutations in Marfan syndrome-related disorders.* Hum Mutat, 2006. **27**: p. 760-9.

105. Miguet, L., et al., *Comparison of a homology model and the crystallographic structure of human 11betahydroxysteroid dehydrogenase type 1 (11beta HSD1) in a structure-based identification of inhibitors.* J Comput Aided Mol Des, 2006. **20**: p. 67-81.

106. Yu, J., et al., *In silico prediction of drug binding to CYP2D6: identification of a new metabolite of metoclopramide.* Drug Metab Dispos, 2006. **34**: p. 1386-92.

107. Prosser, D.E., et al., *Structural Motif-Based Homology Modeling of CYP27A1 and Site-Directed Mutational Analyses Affecting Vitamin D Hydroxylation.* Biophys J, 2006. **90**(10): p. 3389-3409.

108. Wang, Q. and J.R. Halpert, *Combined Three-Dimensional Quantitative Structure-Activity Relationship Analysis of Cytochrome P450 2B6 Substrates and Protein Homology Modeling.* Drug Metab Dispos, 2002. **30**(1): p. 86-95.

109. Lewis, D.F., Y. Ito, and P.S. Goldfarb, *Investigating human P450s involved in drug metabolism via homology with high-resolution P450 crystal structures of the CYP2C subfamily.* Curr Drug Metab, 2006. **7**: p. 589-98.

110. Doniger, S., T. Hofman, and J. Yeh, *Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms.* J Comput Biol, 2002. **9**: p. 849-64.

111. Manallack, D.T. and D.J. Livingstone, *Neural Networks in drug discovery: have theylived up to their promise?* Eur J Med Chem, 1999. **34**: p. 195-208.

112. Yap, C. and Y. Chen, *Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network.* Pharm Sci, 2005. **94**: p. 153-68.

113. Burbidge, R., et al., *Drug design by machine learning: support vector machines for pharmaceutical data analysis.* Comput Chem, 2001. **26**: p. 5-14.

114. Li, H., et al., *Effect of Selection of Molecular Descriptors on the Prediction of Blood-Brain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods.* Journal of Chemical Information and Modeling, 2005. **45**(5): p. 1376-1384.

115. Trotter, M.W.B. and S.B. Holden, *Support Vector machines for ADME property classification.* QSAR Comb Sci, 2003. **22**: p. 533-48.

116. Yap, C.W. and Y.Z. Chen, *Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines.* Journal of Chemical Information and Modeling, 2005. **45**(4): p. 982-992.

117. Binkowski, T.A., A. Joachimiak, and J. Liang, *Protein surface analysis for function annotation in high-throughput structural genomics pipeline.* Protein Sci, 2005. **14**(12): p. 2972-2981.

118. Kang, C.B., et al., *FKBP Family Proteins: Immunophilins with Versatile Biological Functions.* Neurosignals, 2008. **16**: p. 318-25.

119. Wear, M.A. and M.D. Walkinshaw, *Thermodynamics of the cyclophilin-A/cyclosporin-A interaction: a direct comparison of parameters determined by surface plasmon resonance using Biacore T100 and isothermal titration calorimetry.* Anal Biochem, 2006. **359**(2): p. 285-7.

120. Dornan, J., P. Taylor, and M.D. Walkinshaw, *Structures of immunophilins and their ligand complexes.* Curr Top Med Chem, 2003. **3**(12): p. 1392-409.

121. Bramham, J., et al., *Functional Insights from the Structure of the Multifunctional C345C Domain of C5 of Complement.* J Biol Chem, 2005. **280**(11): p. 10636-45.

122. Gasque, P., et al., *Complement components of the innate immune system in health and disease in the CNS.* Immunopharmacology, 2000. **49**(1-2): p. 171-86.

123. Gliubich, F., et al., *Active site structural features for chemically modified forms of rhodanese.* J Biol Chem, 1996. **271**(35): p. 21054-61.

124. Forlani, F., et al., *The cysteine-desulfurase IscS promotes the production of the rhodanese RhdA in the persulfurated form.* FEBS Lett, 2005. **579**(30): p. 6786-90.

125. Yao, H., et al., *An Accurate, Sensitive, and Scalable Method to Identify Functional Sites in Protein Structures.* J Mol Biol, 2003. **326**: p. 255-61.

126. Huang, N., et al., *Molecular mechanics methods for predicting protein–ligand binding.* Phys Chem Chem Phys, 2006. **8**: p. 5166 - 77.

127. Kleywegt, G.J., *Recognition of spatial motifs in protein structures.* J Mol Biol, 1999. **285**: p. 1887-97.

128. Binkowski, T.A., L. Adamian, and J. Liang, *Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns.* J Mol Biol, 2005. **332**: p. 505-26.

129. Jones, S. and J.M. Thornton, *Principles of protein-protein interacions.* Procedings of the National Academy of Science of the United States of America, 1996. **93**(1): p. 13-20.

130. Soga, S., et al., *Identification of the Druggable Concavity in Homology Models Using the PLB Index.* J Chem Inf Model, 2007. **47**(6): p. 2287-2292.

131. Soga, S., et al., *Use of Amino Acid Composition to Predict Ligand-Binding Sites.* J Chem Inf Model, 2007. **47**(2): p. 400-406.

132. Tsai, J., et al., *The packing density in proteins: standard radii and volumes.* J Mol Biol, 1999. **290**(1): p. 253-56.

133. Mehio, W., *A Novel Method to Predict Protein-Protein Binding Surfaces. Masters Thesis*, in *Computer Sciences*. 2007, Chalmers University of Technology: Gothenburgh.

134. Keskin, O., et al., *A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.* Protein Sci, 2004. **13**(4): p. 1043-55.

135. Sayle, R.A. and E.J. Milner-White, *RASMOL: biomolecular graphics for all.* Trends Biochem Sci, 1995. **20**(9): p. 374.

136. Sanner, M.F., A.J. Olson, and J.C. Spehner, *Reduced surface: an efficient way to compute molecular surfaces.* Biopolymers, 1996. **38**(3): p. 305-320.

137. Rispoli, L.A. and T.M. Nett, *Pituitary gonadotropin-releasing hormone (GnRH) receptor: structure, distribution and regulation of expression.* Anim Reprod Sci, 2005. **88**(1-2): p. 57-74.

138. Amory, J.K., S.T. Page, and W.J. Bremner, *Drug Insight: recent advances in male hormonal contraception.* Nat Clin Pract Endorinol Metab, 2006. **2**(1): p. 32-41.

139. Harris, N., et al., *Gonadotropin-releasing hormone gene expression in MDA-MB-231 and ZR-75-1 breast carcinoma cell lines.* Cancer Res, 1991. **51**(10): p. 2577-81.

140. Limonta, P., et al., *Expression of luteinizing hormone-releasing hormone mRNA in the human prostatic cancer cell line LNCaP.* J Clin Endocrinol Metab, 1993. **76**(3): p. 797-800.

141. Pisabarro, M.T., L. Serrano, and M. Wilmanns, *Crystal Structure of the Abl-SH3 Domain Complexed with a Designed High-affinity Peptide Ligand: Implications for SH3-Ligand Interactions.* J Mol Biol, 1998. **281**: p. 513-21.

142. Orlicky, S., et al., *Structural Basis for Phosphodependent Substrate Selection and Orientation by the SCFCdc4 Ubiquitin Ligase.* Cell, 2003. **112**(2): p. 243-56.

143. Cooper, M.S. and P.M. Stewart, *11Beta-Hydroxysteroid Dehydrogenase Type 1 and Its Role in the Hypothalamus-Pituitary-Adrenal Axis, Metabolic*

Syndrome, and Inflammation. J Clin Endocrinol Metab, 2009. **94**(12): p. 4546-54.

144. Laskowski, R.A., et al., *Protein clefts in molecular recognition and function.* Protein Sci, 1996. **5**(12): p. 2438-52.

145. Liang, J., H. Edelsbrunner, and C. Woodward, *Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for drug design.* Prot Sci, 1998. **7**: p. 1884-1897.

146. Weskamp, N., E. Hüllermeier, and G. Klebe, *Merging chemical and biological space: Structural mapping of enzyme binding pocket space.* Proteins, 2009. **76**(2): p. 317-30.

147. Laskowski, R.A., *SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions.* J. Mol. Graph, 1995. **13**: p. 323-30.

148. Huang, B. and M. Schroeder, *LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.* BMC Struct Biol, 2006. **24**: p. 6-19.

149. An, J., M. Totrov, and R. Abagyan, *Pocketome via comprehensive identification and classification of ligand binding envelopes.* Mol Cell Proteomics, 2005. **4**(6): p. 752-61.

150. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* Journal of Molecular Graphics and Modelling, 1997. **15**(6): p. 359-363.

151. Morita, M., S. Nakamura, and K. Shimizu, *Highly acurate method for ligand-binding site prediction in unbound state (apo) protein structures.* Proteins, 2008. **73**(2): p. 468-79.

152. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules.* J Med Chem, 1985. **28**(7): p. 849-57.

153. Laurie, A.T. and R.M. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, 2005. **21**(9): p. 1908-1916.

154. Finney, J.L., *The organization and function of water in protein crystals.* Philos Trans R Soc Lond B Biol Sci, 1977. **278**(959): p. 3-32.

155. Edsall, J.T. and H.A. McKenzie, *Water and proteins. I. The significance and structure of water; its interaction with electrolytes and non-electrolytes.* Adv Biophys, 1978. **10**: p. 137-207.

156. Edsall, J.T. and H.A. McKenzie, *Water and proteins. II. The location and dynamics of water in protein systems and its relation to their stability and properties.* Adv Biophys, 1983. **16**: p. 53-183.

157. Palencia, A., et al., *Role of interfacial water molecules in proline-rich ligand recognition by the Src homology 3 domain of Abl.* J Biol Chem, 2010. **285**(4): p. 2823-33.

158. Di Lella, S., et al., *Critical role of the solvent environment in galectin-1 binding to the disaccharide lactose.* Biochemistry, 2009. **48**(4): p. 786-91.

159. Ladbury, J.E., *Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design.* Chem Biol, 1996. **3**(12): p. 973-80.

160. McDonald, I.K. and J.M. Thornton, *Satisfying Hydrogen Bonding Potential in Proteins.* J Mol Biol, 1994. **238**: p. 777-793.

161. Trott, O. and A.J. Olson, *AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.

162. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.

163. Heyer, L.J., S. Kruglyak, and S. Yooseph, *Exploring expression data: identification and analysis of coexpressed genes.* Genomes Res, 1999. **9**(11): p. 1106-15.

164. Ordonez, C. and E. Omiecinski, *FREM: fast and robust EM clustering for large data sets.* Proceedings of the eleventh international conference on Information and knowledge management, 2002: p. 590-9.

165. Binder, D.A., *Bayesian cluster analysis.* Biometrika, 1978. **65**: p. 31-8.

166. Winters-Hilt, S. and S. Merat, *SVM Clustering.* BMC Bioinformatics, 2007. **8**(Suppl 7): p. S18.

167. Mangiamelia, P., S.K. Chen, and D. West, *A comparison of SOM neural network and hierarchical clustering methods.* Eur J Oper Res, 1996. **93**(2): p. 402-17.

168. Xu, L., *Bayesian Ying–Yang machine, clustering and number of clusters.* Pattern Recognition Letters, 2000. **18**(11-13): p. 1167-78.

169. Emsley, P., et al., *Features and Development of Coot.* Acta Crystallographica Section D - Biological Crystallography, 2010. **66**.

170. Hassan, P., et al., *The CRK3 protein kinase is essential for cell cycle progression of Leishmania mexicana.* Mol Biochem Parasitol, 2001. **113**(2): p. 189-98.

171. Martí-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.

172. Pearson, W. and D. Lipman, *Improved tools for biological sequence comparison.* Proc Natl Acad of Sci U S A, 1988. **85**(8): p. 2444-8.

173. Rost, B., *Twilight zone of protein sequence alignments.* Protein Eng, 1999. **12**(2): p. 85-94.

174. Lobley, A., M.I. Sadowski, and D.T. Jones, *pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination.* Bioinformatics, 2009. **25**(14): p. 1761-7.

175. Lobley, A., et al., *Inferring function using patterns of native disorder in proteins.* PLoS Comput Biol, 2007. **3**(8): p. e162.

176. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

177. Pearson, W.R., *Rapid and sensitive sequence comparison with FASTP and FASTA.* Methods Enzymol, 1990. **183**: p. 63-98.

178. Henikoff, S., J.G. Henikoff, and S. Pietrokovski, *Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations.* Bioinformatics, 1999. **15**(6): p. 471-9.

179. Hoffmann, B., et al., *A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction.* BMC Bioinformatics, 2010. **11**(1): p. 99.

180. Attwood, T.K., *The PRINTS database: a resource for identification of protein families.* Brief Bioinform, 2002. **3**(3): p. 252-63.

181. Falquet, L., et al., *The PROSITE database, its status in 2002.* Nucleic Acids Res, 2002. **30**(1): p. 235-8.

182. Mulder, N.J., et al., *The InterPro Database, 2003 brings increased coverage and new features.* Nucleic Acids Res, 2003. **31**(1): p. 315-8.

183. Park, B.H., et al., *CAZymes Analysis Toolkit (CAT): Web-service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database.* Glycobiology, 2010.

184. Fischer, J.D., G.L. Holliday, and J.M. Thornton, *The CoFactor database: Organic cofactors in enzyme catalysis.* Bioinformation, 2010.

185. Holliday, G.L., J.B. Mitchell, and J.M. Thornton, *Understanding the functional roles of amino acid residues in enzyme catalysis.* J Mol Biol, 2009. **390**(3): p. 560-77.

186. Dobson, P.D. and A.J. Doig, *Predicting Enzyme Class From Protein Structure Without Alignments.* J Mol Biol, 2005. **345**(1): p. 187-99.

187. Frishman, D. and P. Argos, *Knowledge-based protein secondary structure assignment.* Proteins, 1995. **23**(4): p. 566-79.

188. Stawiski, E.W., et al., *Predicting protein function from structure: unique structural features of proteases.* Proc Natl Acad of Sci U S A, 2000. **97**(8): p. 3954-8.

189. Mizuno, M. and B.P. Morgan, *The Possibilities and Pitfalls for Anti-Complement Therapies in Inflammatory Diseases.* Current Drug Traget - Inflammation and Allergy, 2004. **3**(1): p. 87-96.

190. Walport, M., *Complement. First of two parts.* N. Eng. J Med, 2001. **344**(14): p. 1058-66.

191. Walport, M.J., *Complement. Second of two parts.* N. Eng. J Med, 2001. **344**(15): p. 1140-4.

192. Janeway, C.A., et al., *Immunobiology, Sixth edition*. 6 ed. 2005: Garland Science Publishing.

193. Thai, C. and R.T. Ogata, *Recombinant C345C and Factor I Modules of Complement Components C5 and C7 Inhibit C7 Incorporation into the Complement Membrane Attack Complex.* Immunology, 2005. **174**: p. 6227-32.

194. Phelan, M.M., et al., *Solution structure of factor I-like modules from complement C7 reveals a pair of follistatin domains in compact pseudosymmetric arrangement.* J Biol Chem, 2009. **284**(29): p. 19637-49.

195. Stebbins, C.E., W.G.J. Kaelin, and N.P. Pavletich, *Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function.* Science, 1999. **284**(5413): p. 455-61.

196. Gnatt, A.L., et al., *Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution.* Science, 2001. **292**(5523): p. 1876-82.

197. Robinson, R.C., et al., *Crystal structure of Arp2/3 complex.* Science, 2001. **294**(5547): p. 1679-84.

198. Sundberg, E.J., et al., *Structural, energetic, and functional analysis of a protein-protein interface at distinct stages of affinity maturation.* Structure, 2003. **11**(9): p. 1151-61.

199. Harper, J.W., *Protein destruction: Adapting roles for Cks proteins.* Curr Biol, 2001. **11**(11): p. R431 - R435.

200. Sherr, C.J. and J.M. Roberts, *CDK inhibitors: positive and negative regulators of G1-phase progression.* Genes and Development, 1999. **13**(12): p. 1501-12.

201. Tsvetkov, L.M., et al., *p27Kip1 ubiquitination and degradation is regulated by the SCFSkp2 complex through phosphorylated Thr187 in p27.* Curr Biol, 1999. **9**(12): p. 661 - 664, S1-S2.

202. Nguyen, H., D.M. Gitig, and A. Koff, *Cell-Free Degradation of p27kip1, a G1 Cyclin-Dependent Kinase Inhibitor, Is Dependent on CDK2 Activity and the Proteasome.* Mol Cell Biol, 1999. **19**(2): p. 1190-201.

203. Chen, J., et al., *Effects of ectopic overexpression of p21(WAF1/CIP1) on aneuploidy and the malignant phenotype of human brain tumor cells.* Oncogene, 1996. **13**(7): p. 1395-403.

204. Tsihlias, J., L. Kapusta, and J. Slingerland, *The prognostic significance of altered cyclin-dependent kinase inhibitors in human cancer.* Annual Review of Medicine, 1999. **40**(1): p. 401-23.

205. Xu, S., et al., *Substrate Recognition and Ubiquitination of SCFSkp2/Cks1 Ubiquitin-Protein Isopeptide Ligase.* J Biol Chem, 2007. **282**(21): p. 15462-70.

206. Zheng, N., et al., *Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase complex.* Nature, 2002. **416**: p. 703-9.

207. Hao, B., et al., *Structural basis of the Cks1-dependent recognition of p27(Kip1) by the SCF(Skp2) ubiquitin ligase.* Mol Cell, 2005. **20**(1): p. 9-19.

208. Gstaiger, M., et al., *Skp2 is oncogenic and overexpressed in human cancers.* Proc Natl Acad of Sci, 2001. **98**(9): p. 5043-8.

209. Signoretti, S., et al., *Oncogenic role of the ubiquitin ligase subunit Skp2 in human breast cancer.* J Clin Invest, 2002. **110**(5): p. 633-41.

210. Deshaies, R.J., *SCF and cullin/ring H2-based ubiquitin ligases.* Ann Rev Cell Dev Biol, 1999. **15**(1): p. 435-467.

211. Lindsley, J.E. and J. Rutter, *Whence cometh the allosterome.* Proc Natl Acad Sci USA, 2006. **103**(28): p. 10533-5.

212. Tulloch, L.B., et al., *Sulphate removal induces a major conformational change in Leishmania mexicana pyruvate kinase in the crystalline state.* J Mol Biol, 2008. **383**(3): p. 615-26.

213. Morgan, H.P., et al., *An improved strategy for the crystallization of Leishmania mexicana pyruvate kinase.* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2010. **66**(Pt. 3): p. 215-8.

214. Clark M, R. Cramer III, and N. van Opdenbosch, *Validation of the general purpose tripos 5.2 force field.* J Comp Chem, 1989. **10**(8): p. 982-1012.

215. Muller, K., C. Faeh, and F. Diederich, *Fluorine in Pharmaceuticals: Looking Beyond Intuition.* Science, 2007. **317**(5846): p. 1881-6.

216. Dolfing, J. and B.K. Harrison, *Gibbs free energy of formation of halogenated aromatic compounds and their potential role as electron acceptors in anaerobic environments.* Environ Sci Technol, 1992. **26**(11): p. 2213-8.

217. Fersht, A.R., S.E. Jackson, and L. Serrano, *Protein Stability: experimental data from protein engineering.* Philos Trans R Soc Lond A, 1993. **345**: p. 141-151.

218. Kuntz, I., et al., *The maximal affinity of ligands.* Proc Natl Acad of Sci U S A, 1999. **96**(18): p. 9997-10002.

219. Ware, C.F., *The TNF superfamily.* Cytokine Growth Factor Rev, 2003. **14**(3-4): p. 181-4.

220. Hehlgans, T. and K. Pfeffer, *The intriguing biology of the tumour necrosis factor/tumour necrosis factor receptor superfamily: players, rules and the games.* Immunology, 2005. **115**(1): p. 1-20.

221. Eck, M.J. and S.R. Sprang, *The structure of tumor necrosis factor-alpha at 2.6 A resolution. Implications for receptor binding.* J Biol Chem, 1989. **264**(29): p. 17595-605.

222. Loetscher, H., et al., *Human tumor necrosis factor alpha (TNF alpha) mutants with exclusive specificity for the 55-kDa or 75-kDa TNF receptors.* J Biol Chem, 1993. **268**(35): p. 26350-7.

223. Fesik, S.W., *Insights into programmed cell death through structural biology.* Cell, 2000. **103**(2): p. 273-82.

224. Glenney, G.W. and G.D. Wiens, *Early diversification of the TNF superfamily in teleosts: genomic characterization and expression analysis.* J Immunol, 2007. **178**(12): p. 7955-73.

225. Kischkel, F.C., et al., *Apo2L/TRAIL-dependent recruitment of endogenous FADD and caspase-8 to death receptors 4 and 5.* Immunity, 2000. **12**(6): p. 611-20.

226. Darnay, B.G., et al., *Activation of NF-kappaB by RANK requires tumor necrosis factor receptor-associated factor (TRAF) 6 and NF-kappaB-inducing kinase. Identification of a novel TRAF6 interaction motif.* J BIol Chem, 1999. **274**(12): p. 7724-31.

227. Gibson, S.B., et al., *Increased expression of death receptors 4 and 5 synergizes the apoptosis response to combined treatment with etoposide and TRAIL.* Mol Cell Biol, 2000. **20**(1): p. 205-12.

228. Aggarwal, B.B., *Signalling pathways of the TNF superfamily: a double-edged sword.* Nat Rev Immunol, 2003. **3**(9): p. 745-56.

229. Sutherland, A.P., F. Mackay, and C.R. Mackay, *Targeting BAFF: immunomodulation for autoimmune diseases and lymphomas.* Pharmacol Ther, 2006. **112**(3): p. 774-86.

230. Schiemann, B., et al., *An essential role for BAFF in the normal development of B cells through a BCMA-independent pathway.* Science, 2001. **293**(5537): p. 2111-4.

231. Khare, S.D., et al., *Severe B cell hyperplasia and autoimmune disease in TALL-1 transgenic mice.* Proc Natl Acad of Sci U S A, 2000. **97**(7): p. 3370-5.

232. Cheema, G.S., et al., *Elevated serum B lymphocyte stimulator levels in patients with systemic immune-based rheumatic diseases.* Arthritis Rheum, 2001. **44**(6): p. 1313-9.

233. Dillon, S.R., et al., *An APRIL to remember: novel TNF ligands as therapeutic targets.* Nat Rev Drug Discov, 2006. **5**(3): p. 235-46.

234. Tan, S.M., et al., *Local production of B lymphocyte stimulator protein and APRIL in arthritic joints of patients with inflammatory arthritis.* Arthritis Rheum, 2003. **48**(4): p. 982-92.

235. Rennert, P., et al., *A soluble form of B cell maturation antigen, a receptor for the tumor necrosis factor family member APRIL, inhibits tumor cell growth.* J Exp Med, 2000. **192**(11): p. 1677-84.

236. Roodman, G.D. and W.C. Dougall, *RANK ligand as a therapeutic target for bone metastases and multiple myeloma.* Cancer Treat Rev, 2008. **34**(1): p. 92-101.

237. Winkles, J.A., *The TWEAK-Fn14 cytokine-receptor axis: discovery, biology and therapeutic targeting.* Nat Rev Drug Discov, 2008. **7**(5): p. 411-25.

238. Gordon, N.C., et al., *BAFF/BLyS receptor 3 comprises a minimal TNF receptor-like module that encodes a highly focused ligand-binding site.* Biochemistry, 2003. **42**(20): p. 5977-83.

239. Brown, S.A., et al., *TWEAK binding to the Fn14 cysteine-rich domain depends on charged residues located in both the A1 and D2 modules.* Biochem J, 2006. **397**(2): p. 297-304.

240. Fleming, T.J., et al., *Discovery of high-affinity peptide binders to BLyS by phage display.* J Mol Recognit, 2005. **18**(1): p. 94-102.

241. He, M.M., et al., *Small-molecule inhibition of TNF-alpha.* Science, 2005. **310**(5750): p. 1022-5.

242. Gururaja, T.L., et al., *A class of small molecules that inhibit TNFalpha-induced survival and death pathways via prevention of interactions between TNFalphaRI, TRADD, and RIP1.* Chem Biol, 2007. **14**(10): p. 1105-18.

243. Wells, J.A. and C.L. McClendon, *Reaching for high-hanging fruit in drug discovery at protein-protein interfaces.* Nature, 2007. **450**(7172): p. 1001-9.

244. Burkhard, K., et al., *Development of Extracellular Signal-Regulated Kinase Inhibitors.* Curr Top Med Chem, 2009. **9**(8): p. 678-689.

245. Davies, S.P., et al., *Specificity and mechanism of action of some commonly used protein kinase inhibitors.* Biochem J, 2000. **351**(Pt 1): p. 95-105.

246. Bain, J., et al., *The specificities of protein kinase inhibitors: an update.* Biochem J, 2003. **371**(Pt 1): p. 199-204.

247. Fabian, M.A., et al., *A small molecule-kinase interaction map for clinical kinase inhibitors.* Nat Biotechnol, 2005. **23**(3): p. 329-36.

248. Ohori, M., et al., *Identification of a selective ERK inhibitor and structural determination of the inhibitor-ERK2 complex.* Biochem Biophys Res Commun, 2005. **336**(1): p. 357-63.

249. Whitty, A. and G. Kumaravel, *Between a rock and a hard place?* Nat Chem Biol, 2006. **2**(3): p. 112-8.

250. Widmer, A., *WITNOTP: A Computer Program for Molecular Modeling*. 1997, Novartis: Basel.

251. Freidinger, R.M. and D.F. Veber, *Cyclic hexapeptide somatostatin analogs*. 1982, Merck & Co., Inc (Rahway, NJ): United States.

252. Pagliaro, L., et al., *Emerging classes of protein-protein interaction inhibitors and new tools for their development.* Curr Opin Chem Biol, 2004. **8**(4): p. 442-9.

253. Arkin, M.R. and J.A. Wells, *Small-molecule inhibitors of protein-protein interactions: progressing towards the dream.* Nat Rev Drug Discov, 2004. **3**(4): p. 301-17.

254. Archakov, A.I., et al., *Protein-protein interactions as a target for drugs in proteomics.* Proteomics, 2003. **3**(4): p. 380-91.

255. Fry, D.C. and L.T. Vassilev, *Targeting protein-protein interactions for cancer therapy.* J Mol Med, 2005. **83**(12): p. 955-63.

256. Fletcher, S. and A.D. Hamilton, *Targeting protein-protein interactions by rational design: mimicry of protein surfaces.* J R Soc Interface, 2006. **3**(7): p. 215-33.

257. Sharma, S.K., T.M. Ramsey, and K.W. Bair, *Protein-protein interactions: lessons learned.* Curr Med Chem Anticancer Agents, 2002. **2**(2): p. 311-30.

258. Vassilev, L.T., et al., *In vivo activation of the p53 pathway by small-molecule antagonists of MDM2.* Science, 2004. **303**(5659): p. 844-8.

259. Klein, C. and L.T. Vassilev, *Targeting the p53-MDM2 interaction to treat cancer.* Br J Cancer, 2004. **91**(8): p. 1415-9.

260. Gulbis, J.M., et al., *Structure of the C-terminal region of p21(WAF1/CIP1) complexed with human PCNA.* Cell, 1996. **87**(2): p. 297-306.

261. Kontopidis, G., et al., *Structural and biochemical studies of human proliferating cell nuclear antigen complexes provide a rationale for cyclin association and inhibitor design.* Proc Natl Acad of Sci U S A, 2005. **102**(6): p. 1871-6.

262. Warbrick, E., et al., *A small peptide inhibitor of DNA replication defines the site of interaction between the cyclin-dependent kinase inhibitor p21WAF1 and proliferating cell nuclear antigen.* Curr Biol, 1995. **5**(3): p. 275-82.

263. De Benedetti, A. and J.R. Graff, *eIF-4E expression and its role in malignancies and metastases.* Oncogene, 2004. **23**(18): p. 3189-99.

264. Neduva, V. and R.B. Russell, *Linear motifs: evolutionary interaction switches.* FEBS Lett, 2005. **579**(15): p. 3342-5.

265. Puntervoll, P., et al., *ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins.* Nucleic Acids Res, 2003. **31**(13): p. 3625-30.

266. Cochran, A.G., *Protein-protein interfaces: mimics and inhibitors.* Curr Opin Chem Biol, 2001. **5**(6): p. 654-9.

267. Toogood, P.L., *Inhibition of protein-protein association by small molecules: approaches and progress.* J Med Chem, 2002. **45**(8): p. 1543-58.

268. Che, Y., B.R. Brooks, and G.R. Marshall, *Development of small molecules designed to modulate protein-protein interactions.* J Comput Aided Mol Des, 2006. **20**(2): p. 109-30.

269. Li, L., et al., *A small molecule Smac mimic potentiates TRAIL- and TNFalpha-mediated cell death.* Science, 2004. **305**(5689): p. 1471-4.

270. Ding, K., et al., *Structure-based design of potent non-peptide MDM2 inhibitors.* J Am Chem Soc, 2005. **127**(29): p. 10130-1.

271. Tilley, J.W., et al., *Identification of a small molecule inhibitor of the IL-2/IL-2R alpha receptor interaction which binds to IL-2.* J Am Chem Soc, 1997. **119**: p. 7589-90.

272. Berg, T., et al., *Small-molecule antagonists of Myc/Max dimerization inhibit Myc-induced transformation of chicken embryo fibroblasts.* Proc Natl Acad of Sci U S A, 2002. **99**(6): p. 3830-5.

273. Daelemans, D., et al., *A synthetic HIV-1 Rev inhibitor interfering with the CRM1-mediated nuclear export.* Proc Natl Acad of Sci U S A, 2002. **99**(22): p. 14440-5.

274. Oltersdorf, T., et al., *An inhibitor of Bcl-2 family proteins induces regression of solid tumours.* Nature, 2005. **435**(7042): p. 677-81.

275. Degterev, A., et al., *Identification of small-molecule inhibitors of interaction between the BH3 domain and Bcl-xL.* Nat Cell Biol, 2001. **3**(2): p. 173-82.

276. Qureshi, S.A., et al., *Mimicry of erythropoietin by a nonpeptide molecule.* Proc Natl Acad of Sci U S A, 1999. **96**(21): p. 12156-61.

277. Wang, J.L., et al., *Structure-based discovery of an organic compound that binds Bcl-2 protein and induces apoptosis of tumor cells.* Proc Natl Acad of Sci U S A, 2000. **97**(13): p. 7124-9.

278. Kelly, T.A., et al., *Cutting edge: a small molecule antagonist of LFA-1-mediated cell adhesion.* J Immunol, 1999. **163**(10): p. 5173-7.

279. Petsalaki, E. and R.B. Russell, *Peptide-mediated interactions in biological systems: new discoveries and applications.* Curr Opin Biotechnol, 2008. **19**(4): p. 344-50.

280. Vanhee, P., et al., *Protein-peptide interactions adopt the same structural motifs as monomeric protein folds.* Structure, 2009. **17**(8): p. 1128-36.

281. McDonald, I.K., et al., *HBPLUS Hydrogen Bond Calculator*. 1993, Department of Biochemistry and Molecular Biology, University College London.

282. Hubbard, S.J. and J.M. Thornton, *'NACCESS', Computer Program*. 1993, Department of Biochemistry and Molecular Biology, University College London.

283. Taylor, P., *Torsion: A computer program to calculate torsion angles in proteins*. 2003, Institute of Structural and Molecular Biology, University of Edinburgh.

284. Glaser, F., et al., *Residue frequencies and pairing preferences at protein-protein interfaces.* Proteins, 2001. **43**(2): p. 89-102.

285. Ofran, Y. and B. Rost, *Analysing six types of protein-protein interfaces.* J Mol Biol, 2003. **325**(2): p. 377-87.

286. Ansari, S. and V. Helms, *Statistical analysis of predominantly transient protein-protein interfaces.* Proteins, 2005. **61**(2): p. 344-55.

287. Farber, G.K. and G.A. Petsko, *The evolution of alpha/beta barrel enzymes.* Trends Biochem Sci, 1990. **15**(6): p. 228-34.

288. Kannan, N., et al., *Clusters in alpha/beta barrel proteins: implications for protein structure, function, and folding: a graph theoretical approach.* Proteins, 2001. **43**(2): p. 103-12.

289. Pugalenthi, G., et al., *MegaMotifBase: a database of structural motifs in protein families and superfamilies.* Nucleic Acids Res, 2008. **36**(Database Issue): p. D218-21.

290. Todd, A.E., C.A. Orengo, and J.M. Thornton, *Evolution of protein function, from a structural perspective.* Curr Opin Chem Biol, 1999. **3**(5): p. 548-56.

291. Hegyi, H. and M. Gerstein, *The relationship between protein structure and function: a comprehensive survey with application to the yeast genome.* J Mol Biol, 1999. **288**(1): p. 147-64.

292. Schreiter, E.R. and C.L. Drennan, *Ribbon-helix-helix transcription factors: variations on a theme.* Nat Rev Microbiol, 2007. **5**(9): p. 710-20.

293. Marsico, A., et al., *Structural fragment clustering reveals novel structural and functional motifs in alpha-helical transmembrane proteins.* BMC Bioinformatics, 2010. **11**: p. 204.

294. Debret, G., A. Martel, and P. Cuniasse, *RASMOT-3D PRO: a 3D motif search webserver.* Nucleic Acids Res, 2009. **37**(Web Server Issue): p. W459-64.

295. Shulman-Peleg, A., et al., *MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions.* Nucleic Acids Res, 2008. **36**(Web Server Issue): p. W260-4.

296. Goyal, K., D. Mohanty, and S.C. Mande, *PAR-3D: a server to predict protein active site residues.* Nucleic Acids Res, 2007. **35**(Web Server Issue): p. W503-5.

297. Bauer, R.A., et al., *Superimpose: a 3D structural superposition server.* Nucleic Acids Res, 2008. **36**(Web Server Issue): p. W47-54.

298. Madsen, D. and G.T. Kleywegt, *Interactive motif and fold recognition in protein structures.* J Appl Cryst, 2002. **35**: p. 137-9.

299. Velamakanni, S., et al., *A Functional Steroid-Binding Element in an ATP-Binding Cassette Multidrug Transporter.* Mol Pharmacol, 2008. **73**(1): p. 12-7.

300. McFie, P.J., et al., *Identification of a co-repressor that inhibits the transcriptional and growth-arrest activities of CCAAT/enhancer-binding protein alpha.* J BIol Chem, 2006. **281**(26): p. 18069-80.

301. Hofman, K., et al., *The retinoblastoma protein-associated transcription repressor RBaK interacts with the androgen receptor and enhances its transcriptional activity.* J Mol Endocrinol, 2003. **31**(3): p. 583-96.

302. Lee, S., et al., *Structural basis for viral late-domain binding to Alix.* Nat Struct Mol Biol, 2007. **14**(3): p. 194-9.

303. Irie, T., et al., *The YLDL sequence within Sendai virus M protein is critical for budding of virus-like particles and interacts with Alix/AIP1 independently of C protein.* J Virol, 2007. **81**(5): p. 2263-73.

304. Rha, G.B., et al., *Multiple binding modes between HNF4alpha and the LXXLL motifs of PGC-1alpha lead to full activation.* J Biol Chem, 2009. **284**(50): p. 35165-76.

305. Bourgoin-Voillard, S., et al., *Capacity of type I and II ligands to confer to estrogen receptor alpha an appropriate conformation for the recruitment of coactivators containing a LxxLL motif-Relationship with the regulation of receptor level and ERE-dependent transcription in MCF-7 cells.* Biochem Pharmacol, 2010. **79**(5): p. 746-57.

306. Jiang, P., et al., *Key roles for MED1 LxxLL motifs in pubertal mammary gland development and luminal-cell differentiation.* Proc Natl Acad of Sci U S A, 2010. **107**(15): p. 6765-70.

307. Kollara, A. and T.J. Brown, *Variable expression of nuclear receptor coactivator 4 (NcoA4) during mouse embryonic development.* J Histochem Cytochem, 2010. **58**(7): p. 595-609.

308. Heinlein, C.A., et al., *Identification of ARA70 as a ligand-enhanced coactivator for the peroxisome proliferator-activated receptor gamma.* J Biol Chem, 1999. **274**(23): p. 16147-52.

309. Zhou, Z.X., et al., *Domain interactions between coregulator ARA(70) and the androgen receptor (AR).* Mol Endocrinol, 2002. **16**(2): p. 287-300.

310. Ting, H.J., et al., *Androgen-receptor coregulators mediate the suppressive effect of androgen signals on vitamin D receptor activity.* Endocrine, 2005. **26**(1): p. 1-9.

311. Mita, Y., et al., *LXXLL peptide mimetics as inhibitors of the interaction of vitamin D receptor with coactivators.* Bioorg Med Chem Lett, 2010. **20**(5): p. 1712-7.

9 Appendix A: Interface propensity scores for triplets and atomic groups based on the different Score Tables

This appendix presents the calculation results for the different atomic and triangular groups based on each of the Protein-Protein (Table 9-2 and Table 9-5), Protein-Peptide (Table 9-3 and Table 9-6), and Protein-Ligand (Table 9-4 and Table 9-7) score tables.

Table 9-1: Nomenclature of the Atomic Groups

Group	Identifier
C3H0	0
C3H1	1
C4H1	2
C4H2	3
C4H3	4
N3H0	5
N3H1	6
N3H2	7
N4H3	8
O1H0	9
O2H1	A
S2H0	B
S2H1	C

Table 9-2: The propensities and statistics for different atomic triplets in Protein-Protein Binding Sites. ZERO indicates and division by zero result which is not calculatable. In such a case, the triplet is given a score of 0. This work is part of my Masters Degree research. We also present the free energy calculations from these propensities, calculated as ΔG_{stat} = -RTln(2)×propensity.

Type	Interface Count	Surface Count	Total Count	Prob Theoretical	Prob Experimental	Log Likelihood	ΔG_{stat}
000	1	6	7	0.00002	0.000028	-0.497	0.204
001	127	242	369	0.002493	0.001117	1.158	-0.475
002	36	99	135	0.000707	0.000457	0.629	-0.258
003	71	208	279	0.001394	0.00096	0.538	-0.221
004	18	25	43	0.000353	0.000115	1.614	-0.663
005	1	6	7	0.00002	0.000028	-0.497	0.204
006	23	132	155	0.000451	0.000609	-0.433	0.178
007	3	38	41	0.000059	0.000175	-1.575	0.646
008	0	3	3	0	0.000014	ZERO	ZERO

009	68	436	504	0.001335	0.002012	-0.592	0.243
00A	3	16	19	0.000059	0.000074	-0.327	0.134
011	646	629	1275	0.012679	0.002903	2.127	-0.873
012	89	185	274	0.001747	0.000854	1.033	-0.424
013	870	1148	2018	0.017076	0.005299	1.688	-0.693
014	108	167	275	0.00212	0.000771	1.459	-0.599
015	1	1	2	0.00002	0.000005	2.088	-0.857
016	211	510	721	0.004141	0.002354	0.815	-0.335
017	41	107	148	0.000805	0.000494	0.704	-0.289
018	1	6	7	0.00002	0.000028	-0.497	0.204
019	181	457	638	0.003553	0.002109	0.752	-0.309
01A	490	509	999	0.009617	0.002349	2.033	-0.835
01B	3	7	10	0.000059	0.000032	0.866	-0.355
01C	0	1	1	0	0.000005	ZERO	ZERO
022	17	104	121	0.000334	0.00048	-0.525	0.215
023	424	1894	2318	0.008322	0.008742	-0.071	0.029
024	265	625	890	0.005201	0.002885	0.850	-0.349
025	58	195	253	0.001138	0.0009	0.339	-0.139
026	857	3206	4063	0.01682	0.014798	0.185	-0.076
027	55	254	309	0.001079	0.001172	-0.119	0.049
028	2	28	30	0.000039	0.000129	-1.719	0.706
029	1423	6708	8131	0.027929	0.030962	-0.149	0.061
02A	50	307	357	0.000981	0.001417	-0.530	0.218
02B	7	14	21	0.000137	0.000065	1.088	-0.447
02C	8	4	12	0.000157	0.000018	3.088	-1.268
033	249	1584	1833	0.004887	0.007311	-0.581	0.239
034	284	774	1058	0.005574	0.003573	0.642	-0.263
035	21	90	111	0.000412	0.000415	-0.011	0.005
036	889	3603	4492	0.017448	0.01663	0.069	-0.028
037	588	2464	3052	0.011541	0.011373	0.021	-0.009
038	5	119	124	0.000098	0.000549	-2.485	1.020
039	2215	11763	13978	0.043474	0.054294	-0.321	0.132
03A	98	443	541	0.001923	0.002045	-0.088	0.036
03B	17	21	38	0.000334	0.000097	1.783	-0.732
03C	9	2	11	0.000177	0.000009	4.258	-1.748
044	52	116	168	0.001021	0.000535	0.931	-0.382
045	2	12	14	0.000039	0.000055	-0.497	0.204
046	176	487	663	0.003454	0.002248	0.620	-0.254
047	56	180	236	0.001099	0.000831	0.404	-0.166
048	0	7	7	0	0.000032	ZERO	ZERO
049	457	1585	2042	0.00897	0.007316	0.294	-0.121
04A	40	95	135	0.000785	0.000438	0.840	-0.345
04B	6	9	15	0.000118	0.000042	1.503	-0.617
04C	2	1	3	0.000039	0.000005	3.088	-1.268
056	2	26	28	0.000039	0.00012	-1.612	0.662
057	0	6	6	0	0.000028	ZERO	ZERO
059	6	80	86	0.000118	0.000369	-1.649	0.677
05A	0	1	1	0	0.000005	ZERO	ZERO

066	55	211	266	0.001079	0.000974	0.149	-0.061
067	234	837	1071	0.004593	0.003863	0.250	-0.102
068	0	13	13	0	0.00006	ZERO	ZERO
069	400	1981	2381	0.007851	0.009144	-0.220	0.090
06A	26	124	150	0.00051	0.000572	-0.166	0.068
06B	4	4	8	0.000079	0.000018	2.088	-0.857
06C	1	2	3	0.00002	0.000009	1.088	-0.447
077	260	850	1110	0.005103	0.003923	0.379	-0.156
078	1	23	24	0.00002	0.000106	-2.435	1.000
079	508	2239	2747	0.009971	0.010334	-0.052	0.021
07A	16	86	102	0.000314	0.000397	-0.338	0.139
07B	1	4	5	0.00002	0.000018	0.088	-0.036
089	10	251	261	0.000196	0.001159	-2.561	1.051
08A	0	4	4	0	0.000018	ZERO	ZERO
08B	0	5	5	0	0.000023	ZERO	ZERO
099	811	4818	5629	0.015918	0.022238	-0.482	0.198
09A	80	567	647	0.00157	0.002617	-0.737	0.303
09B	12	20	32	0.000236	0.000092	1.351	-0.555
09C	12	8	20	0.000236	0.000037	2.673	-1.097
0AA	2	13	15	0.000039	0.00006	-0.612	0.251
0AB	1	3	4	0.00002	0.000014	0.503	-0.207
0BB	0	1	1	0	0.000005	ZERO	ZERO
111	1403	1192	2595	0.027537	0.005502	2.323	-0.954
112	108	276	384	0.00212	0.001274	0.735	-0.302
113	617	1230	1847	0.01211	0.005677	1.093	-0.449
114	647	799	1446	0.012699	0.003688	1.784	-0.732
116	329	839	1168	0.006457	0.003873	0.738	-0.303
117	136	317	453	0.002669	0.001463	0.867	-0.356
118	11	81	92	0.000216	0.000374	-0.792	0.325
119	418	1055	1473	0.008204	0.004869	0.753	-0.309
11A	157	287	444	0.003081	0.001325	1.218	-0.500
11B	15	29	44	0.000294	0.000134	1.137	-0.467
11C	4	3	7	0.000079	0.000014	2.503	-1.028
122	20	49	69	0.000393	0.000226	0.795	-0.327
123	431	957	1388	0.008459	0.004417	0.937	-0.385
124	255	530	785	0.005005	0.002446	1.033	-0.424
125	0	4	4	0	0.000018	ZERO	ZERO
126	195	611	806	0.003827	0.00282	0.441	-0.181
127	18	59	77	0.000353	0.000272	0.376	-0.154
128	2	18	20	0.000039	0.000083	-1.082	0.444
129	393	1194	1587	0.007713	0.005511	0.485	-0.199
12A	35	163	198	0.000687	0.000752	-0.131	0.054
12B	17	10	27	0.000334	0.000046	2.854	-1.171
12C	2	2	4	0.000039	0.000009	2.088	-0.857
133	349	1238	1587	0.00685	0.005714	0.262	-0.107
134	526	1003	1529	0.010324	0.004629	1.157	-0.475
135	0	4	4	0	0.000018	ZERO	ZERO
136	345	1321	1666	0.006771	0.006097	0.151	-0.062

137	91	451	542	0.001786	0.002082	-0.221	0.091
138	24	226	250	0.000471	0.001043	-1.147	0.471
139	622	2428	3050	0.012208	0.011207	0.123	-0.051
13A	180	625	805	0.003533	0.002885	0.292	-0.120
13B	35	60	95	0.000687	0.000277	1.311	-0.538
13C	10	18	28	0.000196	0.000083	1.240	-0.509
144	430	590	1020	0.00844	0.002723	1.632	-0.670
146	149	504	653	0.002924	0.002326	0.330	-0.136
147	38	142	180	0.000746	0.000655	0.186	-0.077
148	2	21	23	0.000039	0.000097	-1.304	0.535
149	196	701	897	0.003847	0.003236	0.250	-0.102
14A	143	409	552	0.002807	0.001888	0.572	-0.235
14B	28	60	88	0.00055	0.000277	0.989	-0.406
14C	3	22	25	0.000059	0.000102	-0.786	0.323
156	0	1	1	0	0.000005	ZERO	ZERO
159	1	2	3	0.00002	0.000009	1.088	-0.447
15B	0	1	1	0	0.000005	ZERO	ZERO
166	21	151	172	0.000412	0.000697	-0.758	0.311
167	34	185	219	0.000667	0.000854	-0.356	0.146
168	2	74	76	0.000039	0.000342	-3.121	1.281
169	264	1075	1339	0.005182	0.004962	0.063	-0.026
16A	40	177	217	0.000785	0.000817	-0.057	0.024
16B	7	13	20	0.000137	0.00006	1.195	-0.491
16C	10	1	11	0.000196	0.000005	5.410	-2.221
177	24	131	155	0.000471	0.000605	-0.360	0.148
178	1	10	11	0.00002	0.000046	-1.234	0.506
179	80	390	470	0.00157	0.0018	-0.197	0.081
17A	28	134	162	0.00055	0.000618	-0.170	0.070
17B	0	2	2	0	0.000009	ZERO	ZERO
17C	0	1	1	0	0.000005	ZERO	ZERO
189	3	100	103	0.000059	0.000462	-2.971	1.219
18A	4	32	36	0.000079	0.000148	-0.912	0.374
18B	1	3	4	0.00002	0.000014	0.503	-0.207
199	85	459	544	0.001668	0.002119	-0.345	0.141
19A	90	373	463	0.001766	0.001722	0.037	-0.015
19B	17	12	29	0.000334	0.000055	2.591	-1.063
19C	3	5	8	0.000059	0.000023	1.351	-0.555
1AA	31	71	102	0.000608	0.000328	0.893	-0.366
1AB	4	12	16	0.000079	0.000055	0.503	-0.207
1AC	1	0	1	0.00002	0	ZERO	ZERO
1BB	0	2	2	0	0.000009	ZERO	ZERO
222	6	7	13	0.000118	0.000032	1.866	-0.766
223	112	304	416	0.002198	0.001403	0.648	-0.266
224	368	934	1302	0.007223	0.004311	0.745	-0.306
225	1	4	5	0.00002	0.000018	0.088	-0.036
226	137	413	550	0.002689	0.001906	0.496	-0.204
227	4	40	44	0.000079	0.000185	-1.234	0.506
228	1	8	9	0.00002	0.000037	-0.912	0.374

229	206	852	1058	0.004043	0.003933	0.040	-0.016
22A	51	198	249	0.001001	0.000914	0.131	-0.054
22B	1	2	3	0.00002	0.000009	1.088	-0.447
22C	1	15	16	0.00002	0.000069	-1.819	0.747
233	706	3794	4500	0.013857	0.017512	-0.338	0.139
234	1290	3551	4841	0.025319	0.01639	0.627	-0.258
235	76	324	400	0.001492	0.001495	-0.004	0.002
236	1238	5370	6608	0.024298	0.024786	-0.029	0.012
237	172	862	1034	0.003376	0.003979	-0.237	0.097
238	9	253	262	0.000177	0.001168	-2.725	1.118
239	2074	11027	13101	0.040707	0.050897	-0.322	0.132
23A	207	1023	1230	0.004063	0.004722	-0.217	0.089
23B	41	91	132	0.000805	0.00042	0.938	-0.385
23C	47	40	87	0.000922	0.000185	2.321	-0.953
244	1267	2356	3623	0.024868	0.010874	1.193	-0.490
245	2	19	21	0.000039	0.000088	-1.160	0.476
246	802	2607	3409	0.015741	0.012033	0.388	-0.159
247	68	412	480	0.001335	0.001902	-0.511	0.210
248	5	96	101	0.000098	0.000443	-2.175	0.893
249	1326	5493	6819	0.026026	0.025354	0.038	-0.015
24A	315	1365	1680	0.006183	0.0063	-0.027	0.011
24B	40	49	89	0.000785	0.000226	1.795	-0.737
24C	16	47	63	0.000314	0.000217	0.534	-0.219
256	0	8	8	0	0.000037	ZERO	ZERO
257	0	12	12	0	0.000055	ZERO	ZERO
259	10	62	72	0.000196	0.000286	-0.544	0.223
25A	0	9	9	0	0.000042	ZERO	ZERO
266	141	573	714	0.002767	0.002645	0.065	-0.027
267	63	388	451	0.001237	0.001791	-0.534	0.219
268	1	90	91	0.00002	0.000415	-4.404	1.808
269	938	4613	5551	0.01841	0.021292	-0.210	0.086
26A	174	827	1001	0.003415	0.003817	-0.161	0.066
26B	12	41	53	0.000236	0.000189	0.316	-0.130
26C	13	37	50	0.000255	0.000171	0.579	-0.238
277	6	56	62	0.000118	0.000258	-1.134	0.466
278	0	11	11	0	0.000051	ZERO	ZERO
279	139	944	1083	0.002728	0.004357	-0.675	0.277
27A	12	110	122	0.000236	0.000508	-1.108	0.455
27B	6	5	11	0.000118	0.000023	2.351	-0.965
27C	1	1	2	0.00002	0.000005	2.088	-0.857
289	4	288	292	0.000079	0.001329	-4.082	1.675
28A	0	18	18	0	0.000083	ZERO	ZERO
299	453	3537	3990	0.008891	0.016325	-0.877	0.360
29A	158	1105	1263	0.003101	0.0051	-0.718	0.295
29B	26	80	106	0.00051	0.000369	0.467	-0.192
29C	32	26	58	0.000628	0.00012	2.388	-0.980
2AA	16	94	110	0.000314	0.000434	-0.466	0.191
2AB	0	3	3	0	0.000014	ZERO	ZERO

2AC	1	3	4	0.00002	0.000014	0.503	-0.207
2BB	0	3	3	0	0.000014	ZERO	ZERO
2CC	0	1	1	0	0.000005	ZERO	ZERO
333	606	4534	5140	0.011894	0.020927	-0.815	0.335
334	733	2698	3431	0.014387	0.012453	0.208	-0.085
335	35	157	192	0.000687	0.000725	-0.077	0.032
336	611	3383	3994	0.011992	0.015615	-0.381	0.156
337	289	1723	2012	0.005672	0.007953	-0.488	0.200
338	251	2203	2454	0.004926	0.010168	-1.045	0.429
339	1255	10010	11265	0.024632	0.046202	-0.907	0.372
33A	128	1126	1254	0.002512	0.005197	-1.049	0.430
33B	78	236	314	0.001531	0.001089	0.491	-0.202
33C	13	36	49	0.000255	0.000166	0.619	-0.254
344	1259	2456	3715	0.024711	0.011336	1.124	-0.461
345	5	24	29	0.000098	0.000111	-0.175	0.072
346	471	1846	2317	0.009244	0.00852	0.118	-0.048
347	172	940	1112	0.003376	0.004339	-0.362	0.149
348	53	491	544	0.00104	0.002266	-1.123	0.461
349	1369	5881	7250	0.026869	0.027145	-0.015	0.006
34A	168	966	1134	0.003297	0.004459	-0.435	0.179
34B	192	350	542	0.003768	0.001615	1.222	-0.502
34C	25	59	84	0.000491	0.000272	0.849	-0.349
356	3	26	29	0.000059	0.00012	-1.027	0.422
357	0	5	5	0	0.000023	ZERO	ZERO
359	16	123	139	0.000314	0.000568	-0.854	0.351
35A	0	7	7	0	0.000032	ZERO	ZERO
35B	0	2	2	0	0.000009	ZERO	ZERO
366	175	1208	1383	0.003435	0.005576	-0.699	0.287
367	140	929	1069	0.002748	0.004288	-0.642	0.264
368	7	143	150	0.000137	0.00066	-2.264	0.929
369	1580	8063	9643	0.031011	0.037216	-0.263	0.108
36A	196	1089	1285	0.003847	0.005026	-0.386	0.158
36B	30	111	141	0.000589	0.000512	0.201	-0.082
36C	14	24	38	0.000275	0.000111	1.311	-0.538
377	71	451	522	0.001394	0.002082	-0.579	0.238
378	22	231	253	0.000432	0.001066	-1.304	0.535
379	361	3413	3774	0.007085	0.015753	-1.153	0.473
37A	72	383	455	0.001413	0.001768	-0.323	0.133
37B	11	53	64	0.000216	0.000245	-0.180	0.074
37C	3	6	9	0.000059	0.000028	1.088	-0.447
388	2	67	69	0.000039	0.000309	-2.978	1.222
389	89	1777	1866	0.001747	0.008202	-2.231	0.916
38A	31	211	242	0.000608	0.000974	-0.679	0.279
38B	0	17	17	0	0.000078	ZERO	ZERO
38C	2	0	2	0.000039	0	ZERO	ZERO
399	653	5762	6415	0.012816	0.026595	-1.053	0.432
39A	300	2176	2476	0.005888	0.010044	-0.770	0.316
39B	62	221	283	0.001217	0.00102	0.255	-0.104

39C	45	65	110	0.000883	0.0003	1.558	-0.639
3AA	19	183	202	0.000373	0.000845	-1.180	0.484
3AB	8	31	39	0.000157	0.000143	0.134	-0.055
3AC	1	6	7	0.00002	0.000028	-0.497	0.204
3BB	8	19	27	0.000157	0.000088	0.840	-0.345
3CC	2	0	2	0.000039	0	ZERO	ZERO
444	676	992	1668	0.013268	0.004579	1.535	-0.630
445	0	1	1	0	0.000005	ZERO	ZERO
446	92	420	512	0.001806	0.001939	-0.102	0.042
447	87	312	399	0.001708	0.00144	0.246	-0.101
448	7	60	67	0.000137	0.000277	-1.011	0.415
449	397	1632	2029	0.007792	0.007533	0.049	-0.020
44A	99	410	509	0.001943	0.001892	0.038	-0.016
44B	72	119	191	0.001413	0.000549	1.363	-0.560
44C	8	21	29	0.000157	0.000097	0.696	-0.286
456	1	2	3	0.00002	0.000009	1.088	-0.447
459	0	2	2	0	0.000009	ZERO	ZERO
45A	0	2	2	0	0.000009	ZERO	ZERO
466	45	241	286	0.000883	0.001112	-0.333	0.137
467	41	240	281	0.000805	0.001108	-0.461	0.189
468	0	24	24	0	0.000111	ZERO	ZERO
469	482	2118	2600	0.00946	0.009776	-0.047	0.019
46A	74	331	405	0.001452	0.001528	-0.073	0.030
46B	2	20	22	0.000039	0.000092	-1.234	0.506
46C	2	7	9	0.000039	0.000032	0.281	-0.115
477	37	192	229	0.000726	0.000886	-0.287	0.118
478	0	32	32	0	0.000148	ZERO	ZERO
479	128	1073	1201	0.002512	0.004953	-0.979	0.402
47A	19	186	205	0.000373	0.000859	-1.203	0.494
47B	3	22	25	0.000059	0.000102	-0.786	0.323
47C	1	7	8	0.00002	0.000032	-0.719	0.295
488	0	2	2	0	0.000009	ZERO	ZERO
489	17	251	268	0.000334	0.001159	-1.796	0.737
48A	0	55	55	0	0.000254	ZERO	ZERO
48B	1	1	2	0.00002	0.000005	2.088	-0.857
499	160	1463	1623	0.00314	0.006753	-1.105	0.453
49A	110	836	946	0.002159	0.003859	-0.838	0.344
49B	30	102	132	0.000589	0.000471	0.323	-0.132
49C	8	18	26	0.000157	0.000083	0.918	-0.377
4AA	15	96	111	0.000294	0.000443	-0.590	0.242
4AB	6	27	33	0.000118	0.000125	-0.082	0.034
4AC	0	5	5	0	0.000023	ZERO	ZERO
4BB	11	8	19	0.000216	0.000037	2.548	-1.046
4BC	4	2	6	0.000079	0.000009	3.088	-1.268
4CC	0	1	1	0	0.000005	ZERO	ZERO
566	0	1	1	0	0.000005	ZERO	ZERO
567	0	2	2	0	0.000009	ZERO	ZERO
569	0	5	5	0	0.000023	ZERO	ZERO

56B	0	1	1	0	0.000005	ZERO	ZERO
579	0	4	4	0	0.000018	ZERO	ZERO
57A	0	1	1	0	0.000005	ZERO	ZERO
599	3	9	12	0.000059	0.000042	0.503	-0.207
59A	0	2	2	0	0.000009	ZERO	ZERO
5AA	0	2	2	0	0.000009	ZERO	ZERO
666	1	42	43	0.00002	0.000194	-3.304	1.356
667	14	45	59	0.000275	0.000208	0.404	-0.166
668	0	4	4	0	0.000018	ZERO	ZERO
669	84	557	641	0.001649	0.002571	-0.641	0.263
66A	20	143	163	0.000393	0.00066	-0.750	0.308
66B	1	6	7	0.00002	0.000028	-0.497	0.204
66C	1	2	3	0.00002	0.000009	1.088	-0.447
677	9	88	97	0.000177	0.000406	-1.201	0.493
678	0	11	11	0	0.000051	ZERO	ZERO
679	80	722	802	0.00157	0.003332	-1.086	0.446
67A	16	82	98	0.000314	0.000378	-0.269	0.111
67B	1	12	13	0.00002	0.000055	-1.497	0.614
689	2	67	69	0.000039	0.000309	-2.978	1.222
68A	1	5	6	0.00002	0.000023	-0.234	0.096
699	165	1367	1532	0.003238	0.00631	-0.962	0.395
69A	110	628	738	0.002159	0.002899	-0.425	0.174
69B	7	35	42	0.000137	0.000162	-0.234	0.096
69C	5	19	24	0.000098	0.000088	0.162	-0.067
6AA	6	46	52	0.000118	0.000212	-0.850	0.349
6AB	1	3	4	0.00002	0.000014	0.503	-0.207
6BB	0	1	1	0	0.000005	ZERO	ZERO
777	8	87	95	0.000157	0.000402	-1.355	0.556
778	1	17	18	0.00002	0.000078	-1.999	0.821
779	91	604	695	0.001786	0.002788	-0.642	0.264
77A	6	65	71	0.000118	0.0003	-1.349	0.554
77B	0	10	10	0	0.000046	ZERO	ZERO
77C	0	2	2	0	0.000009	ZERO	ZERO
788	0	1	1	0	0.000005	ZERO	ZERO
789	9	143	152	0.000177	0.00066	-1.902	0.781
78A	3	6	9	0.000059	0.000028	1.088	-0.447
799	92	1099	1191	0.001806	0.005073	-1.490	0.612
79A	45	291	336	0.000883	0.001343	-0.605	0.248
79B	9	26	35	0.000177	0.00012	0.558	-0.229
79C	4	10	14	0.000079	0.000046	0.766	-0.315
7AA	5	20	25	0.000098	0.000092	0.088	-0.036
7AB	0	1	1	0	0.000005	ZERO	ZERO
7BB	0	1	1	0	0.000005	ZERO	ZERO
889	0	29	29	0	0.000134	ZERO	ZERO
899	15	367	382	0.000294	0.001694	-2.525	1.036
89A	6	80	86	0.000118	0.000369	-1.649	0.677
89B	0	10	10	0	0.000046	ZERO	ZERO
8AA	0	8	8	0	0.000037	ZERO	ZERO

8AB	0	2	2	0	0.000009	ZERO	ZERO
999	78	666	744	0.001531	0.003074	-1.006	0.413
99A	48	507	555	0.000942	0.00234	-1.313	0.539
99B	11	21	32	0.000216	0.000097	1.155	-0.474
99C	3	9	12	0.000059	0.000042	0.503	-0.207
9AA	17	115	132	0.000334	0.000531	-0.670	0.275
9AB	1	9	10	0.00002	0.000042	-1.082	0.444
9AC	2	6	8	0.000039	0.000028	0.503	-0.207
9BB	1	13	14	0.00002	0.00006	-1.612	0.662
AAA	1	9	10	0.00002	0.000042	-1.082	0.444

Table 9-3: The propensities and statistics for different atomic triplets in Protein-Peptide Binding Sites. ZERO indicates and division by zero result which is not calculatable. In such a case, the triplet is given a score of 0. We also present the free energy calculations from these propensities, calculated as ΔG_{stat} = -RTln(2)×propensity.

Type	Interface Count	Surface Count	Total Count	Prob Theoretical	Prob Experimental	Log Likelihood	ΔG_{stat}
000	0	22	22	0	0.000016	ZERO	ZERO
001	286	1449	1735	0.004836	0.001023	2.241	-0.920
002	34	668	702	0.000575	0.000472	0.286	-0.117
003	97	1108	1205	0.00164	0.000782	1.068	-0.438
004	28	185	213	0.000473	0.000131	1.858	-0.763
005	1	77	78	0.000017	0.000054	-1.685	0.692
006	52	935	987	0.000879	0.00066	0.414	-0.170
007	6	185	191	0.000101	0.000131	-0.365	0.150
008	0	25	25	0	0.000018	ZERO	ZERO
009	119	2931	3050	0.002012	0.002069	-0.040	0.016
00A	20	93	113	0.000338	0.000066	2.365	-0.971
00B	2	7	9	0.000034	0.000005	2.775	-1.139
00C	0	3	3	0	0.000002	ZERO	ZERO
011	711	5200	5911	0.012021	0.003671	1.711	-0.702
012	100	1159	1259	0.001691	0.000818	1.047	-0.430
013	849	8167	9016	0.014355	0.005766	1.316	-0.540
014	174	1095	1269	0.002942	0.000773	1.928	-0.791
015	0	15	15	0	0.000011	ZERO	ZERO
016	392	3261	3653	0.006628	0.002302	1.526	-0.626
017	73	754	827	0.001234	0.000532	1.213	-0.498
018	3	75	78	0.000051	0.000053	-0.062	0.025
019	251	3337	3588	0.004244	0.002356	0.849	-0.348
01A	542	4727	5269	0.009164	0.003337	1.457	-0.598
01B	9	29	38	0.000152	0.00002	2.894	-1.188
01C	5	23	28	0.000085	0.000016	2.380	-0.977
022	19	635	654	0.000321	0.000448	-0.481	0.197
023	379	12121	12500	0.006408	0.008557	-0.417	0.171
024	221	3828	4049	0.003737	0.002703	0.467	-0.192

025	29	1296	1325	0.00049	0.000915	-0.900	0.369
026	794	19704	20498	0.013425	0.013911	-0.051	0.021
027	48	1860	1908	0.000812	0.001313	-0.694	0.285
028	0	210	210	0	0.000148	ZERO	ZERO
029	1276	43178	44454	0.021574	0.030483	-0.499	0.205
02A	52	2075	2127	0.000879	0.001465	-0.737	0.303
02B	4	43	47	0.000068	0.00003	1.156	-0.475
02C	9	106	115	0.000152	0.000075	1.024	-0.420
033	326	10282	10608	0.005512	0.007259	-0.397	0.163
034	295	4628	4923	0.004988	0.003267	0.610	-0.250
035	20	656	676	0.000338	0.000463	-0.454	0.186
036	866	23590	24456	0.014642	0.016654	-0.186	0.076
037	534	17035	17569	0.009029	0.012026	-0.414	0.170
038	10	784	794	0.000169	0.000553	-1.711	0.702
039	2096	77563	79659	0.035438	0.054758	-0.628	0.258
03A	114	2926	3040	0.001927	0.002066	-0.100	0.041
03B	17	149	166	0.000287	0.000105	1.450	-0.595
03C	7	137	144	0.000118	0.000097	0.291	-0.119
044	62	666	728	0.001048	0.00047	1.157	-0.475
045	3	56	59	0.000051	0.00004	0.359	-0.147
046	164	2674	2838	0.002773	0.001888	0.555	-0.228
047	66	1122	1188	0.001116	0.000792	0.494	-0.203
048	1	37	38	0.000017	0.000026	-0.628	0.258
049	428	8909	9337	0.007236	0.00629	0.202	-0.083
04A	47	647	694	0.000795	0.000457	0.799	-0.328
04B	7	44	51	0.000118	0.000031	1.930	-0.792
04C	1	28	29	0.000017	0.00002	-0.225	0.092
055	0	1	1	0	0.000001	ZERO	ZERO
056	2	150	152	0.000034	0.000106	-1.647	0.676
057	0	16	16	0	0.000011	ZERO	ZERO
058	0	4	4	0	0.000003	ZERO	ZERO
059	10	505	515	0.000169	0.000357	-1.076	0.442
05A	1	26	27	0.000017	0.000018	-0.119	0.049
05C	0	8	8	0	0.000006	ZERO	ZERO
066	32	1199	1231	0.000541	0.000846	-0.646	0.265
067	229	6428	6657	0.003872	0.004538	-0.229	0.094
068	4	125	129	0.000068	0.000088	-0.384	0.158
069	506	14138	14644	0.008555	0.009981	-0.222	0.091
06A	39	907	946	0.000659	0.00064	0.042	-0.017
06B	4	31	35	0.000068	0.000022	1.628	-0.668
06C	3	38	41	0.000051	0.000027	0.919	-0.377
077	240	6569	6809	0.004058	0.004638	-0.193	0.079
078	2	133	135	0.000034	0.000094	-1.473	0.605
079	459	15906	16365	0.007761	0.011229	-0.533	0.219
07A	30	531	561	0.000507	0.000375	0.436	-0.179
07B	4	29	33	0.000068	0.00002	1.724	-0.708
07C	2	20	22	0.000034	0.000014	1.260	-0.517
088	0	1	1	0	0.000001	ZERO	ZERO

089	19	1484	1503	0.000321	0.001048	-1.705	0.700
08A	3	52	55	0.000051	0.000037	0.466	-0.191
099	917	32872	33789	0.015504	0.023207	-0.582	0.239
09A	145	3913	4058	0.002452	0.002763	-0.172	0.071
09B	13	163	176	0.00022	0.000115	0.934	-0.383
09C	18	184	202	0.000304	0.00013	1.228	-0.504
0AA	13	129	142	0.00022	0.000091	1.271	-0.522
0AB	1	8	9	0.000017	0.000006	1.582	-0.649
0AC	1	6	7	0.000017	0.000004	1.997	-0.820
0BB	0	2	2	0	0.000001	ZERO	ZERO
111	1570	10116	11686	0.026545	0.007142	1.894	-0.777
112	164	2073	2237	0.002773	0.001464	0.922	-0.378
113	822	9231	10053	0.013898	0.006517	1.093	-0.449
114	1071	6690	7761	0.018108	0.004723	1.939	-0.796
115	0	6	6	0	0.000004	ZERO	ZERO
116	455	5563	6018	0.007693	0.003927	0.970	-0.398
117	199	2081	2280	0.003365	0.001469	1.195	-0.491
118	14	493	507	0.000237	0.000348	-0.556	0.228
119	731	8260	8991	0.012359	0.005831	1.084	-0.445
11A	282	2103	2385	0.004768	0.001485	1.683	-0.691
11B	39	214	253	0.000659	0.000151	2.126	-0.873
11C	15	134	149	0.000254	0.000095	1.423	-0.584
122	24	415	439	0.000406	0.000293	0.470	-0.193
123	447	7385	7832	0.007558	0.005214	0.536	-0.220
124	333	3800	4133	0.00563	0.002683	1.069	-0.439
125	0	33	33	0	0.000023	ZERO	ZERO
126	277	4362	4639	0.004683	0.00308	0.605	-0.248
127	28	519	547	0.000473	0.000366	0.370	-0.152
128	4	104	108	0.000068	0.000073	-0.119	0.049
129	508	8528	9036	0.008589	0.006021	0.513	-0.211
12A	73	1275	1348	0.001234	0.0009	0.455	-0.187
12B	5	61	66	0.000085	0.000043	0.973	-0.399
12C	12	94	106	0.000203	0.000066	1.612	-0.662
133	513	8715	9228	0.008674	0.006153	0.495	-0.203
134	844	7627	8471	0.01427	0.005385	1.406	-0.577
135	2	25	27	0.000034	0.000018	0.938	-0.385
136	506	9628	10134	0.008555	0.006797	0.332	-0.136
137	168	2953	3121	0.00284	0.002085	0.446	-0.183
138	37	1435	1472	0.000626	0.001013	-0.695	0.285
139	844	16731	17575	0.01427	0.011812	0.273	-0.112
13A	277	5037	5314	0.004683	0.003556	0.397	-0.163
13B	39	386	425	0.000659	0.000273	1.275	-0.523
13C	33	286	319	0.000558	0.000202	1.466	-0.602
144	760	4621	5381	0.01285	0.003262	1.978	-0.812
145	0	2	2	0	0.000001	ZERO	ZERO
146	240	3350	3590	0.004058	0.002365	0.779	-0.320
147	78	1006	1084	0.001319	0.00071	0.893	-0.367
148	6	181	187	0.000101	0.000128	-0.333	0.137

149	352	5107	5459	0.005951	0.003605	0.723	-0.297
14A	308	3261	3569	0.005208	0.002302	1.178	-0.484
14B	63	379	442	0.001065	0.000268	1.993	-0.818
14C	21	116	137	0.000355	0.000082	2.116	-0.869
156	0	10	10	0	0.000007	ZERO	ZERO
158	0	1	1	0	0.000001	ZERO	ZERO
159	2	14	16	0.000034	0.00001	1.775	-0.729
15A	0	6	6	0	0.000004	ZERO	ZERO
166	60	976	1036	0.001014	0.000689	0.558	-0.229
167	80	1592	1672	0.001353	0.001124	0.267	-0.110
168	11	256	267	0.000186	0.000181	0.041	-0.017
169	462	7336	7798	0.007811	0.005179	0.593	-0.243
16A	83	1311	1394	0.001403	0.000926	0.600	-0.246
16B	7	91	98	0.000118	0.000064	0.881	-0.362
16C	17	103	120	0.000287	0.000073	1.983	-0.814
177	58	732	790	0.000981	0.000517	0.924	-0.379
178	3	44	47	0.000051	0.000031	0.707	-0.290
179	148	2672	2820	0.002502	0.001886	0.408	-0.167
17A	62	1031	1093	0.001048	0.000728	0.526	-0.216
17B	5	30	35	0.000085	0.000021	1.997	-0.820
17C	2	23	25	0.000034	0.000016	1.058	-0.434
188	1	8	9	0.000017	0.000006	1.582	-0.649
189	14	524	538	0.000237	0.00037	-0.644	0.264
18A	5	272	277	0.000085	0.000192	-1.184	0.486
18B	0	3	3	0	0.000002	ZERO	ZERO
18C	0	5	5	0	0.000004	ZERO	ZERO
199	162	3255	3417	0.002739	0.002298	0.253	-0.104
19A	204	3030	3234	0.003449	0.002139	0.689	-0.283
19B	14	116	130	0.000237	0.000082	1.531	-0.628
19C	7	103	110	0.000118	0.000073	0.703	-0.289
1AA	97	511	608	0.00164	0.000361	2.185	-0.897
1AB	13	65	78	0.00022	0.000046	2.260	-0.928
1AC	4	57	61	0.000068	0.00004	0.749	-0.307
1BB	1	11	12	0.000017	0.000008	1.122	-0.461
1BC	0	4	4	0	0.000003	ZERO	ZERO
1CC	1	3	4	0.000017	0.000002	2.997	-1.230
222	5	142	147	0.000085	0.0001	-0.246	0.101
223	81	2303	2384	0.00137	0.001626	-0.248	0.102
224	322	5852	6174	0.005444	0.004131	0.398	-0.163
225	1	34	35	0.000017	0.000024	-0.506	0.208
226	106	2866	2972	0.001792	0.002023	-0.175	0.072
227	11	272	283	0.000186	0.000192	-0.046	0.019
228	0	49	49	0	0.000035	ZERO	ZERO
229	224	5748	5972	0.003787	0.004058	-0.100	0.041
22A	52	1249	1301	0.000879	0.000882	-0.004	0.002
22B	0	18	18	0	0.000013	ZERO	ZERO
22C	0	35	35	0	0.000025	ZERO	ZERO
233	611	26347	26958	0.010331	0.018601	-0.848	0.348

234	1207	22536	23743	0.020407	0.01591	0.359	-0.147
235	56	2194	2250	0.000947	0.001549	-0.710	0.291
236	1124	35323	36447	0.019004	0.024938	-0.392	0.161
237	139	5748	5887	0.00235	0.004058	-0.788	0.323
238	21	1722	1743	0.000355	0.001216	-1.776	0.729
239	1845	72864	74709	0.031195	0.051441	-0.722	0.296
23A	216	6812	7028	0.003652	0.004809	-0.397	0.163
23B	51	637	688	0.000862	0.00045	0.939	-0.385
23C	45	624	669	0.000761	0.000441	0.788	-0.323
244	1186	14023	15209	0.020052	0.0099	1.018	-0.418
245	4	98	102	0.000068	0.000069	-0.033	0.014
246	825	15458	16283	0.013949	0.010913	0.354	-0.145
247	74	2713	2787	0.001251	0.001915	-0.614	0.252
248	13	542	555	0.00022	0.000383	-0.800	0.328
249	1421	32370	33791	0.024026	0.022853	0.072	-0.030
24A	343	8840	9183	0.005799	0.006241	-0.106	0.044
24B	47	424	471	0.000795	0.000299	1.409	-0.578
24C	25	231	256	0.000423	0.000163	1.374	-0.564
255	0	1	1	0	0.000001	ZERO	ZERO
256	0	56	56	0	0.00004	ZERO	ZERO
257	0	36	36	0	0.000025	ZERO	ZERO
258	0	6	6	0	0.000004	ZERO	ZERO
259	9	500	509	0.000152	0.000353	-1.214	0.498
25A	1	32	33	0.000017	0.000023	-0.418	0.172
25B	0	1	1	0	0.000001	ZERO	ZERO
25C	0	4	4	0	0.000003	ZERO	ZERO
266	95	3842	3937	0.001606	0.002712	-0.756	0.310
267	69	2723	2792	0.001167	0.001922	-0.721	0.296
268	10	471	481	0.000169	0.000333	-0.976	0.401
269	1016	29569	30585	0.017178	0.020875	-0.281	0.115
26A	149	4891	5040	0.002519	0.003453	-0.455	0.187
26B	12	223	235	0.000203	0.000157	0.366	-0.150
26C	25	292	317	0.000423	0.000206	1.036	-0.425
277	7	425	432	0.000118	0.0003	-1.342	0.551
278	0	66	66	0	0.000047	ZERO	ZERO
279	134	6912	7046	0.002266	0.00488	-1.107	0.454
27A	26	688	714	0.00044	0.000486	-0.144	0.059
27B	4	31	35	0.000068	0.000022	1.628	-0.668
27C	1	38	39	0.000017	0.000027	-0.666	0.273
288	0	4	4	0	0.000003	ZERO	ZERO
289	25	1886	1911	0.000423	0.001331	-1.655	0.679
28A	6	159	165	0.000101	0.000112	-0.146	0.060
28B	0	2	2	0	0.000001	ZERO	ZERO
28C	0	6	6	0	0.000004	ZERO	ZERO
299	588	22759	23347	0.009942	0.016068	-0.693	0.284
29A	222	6976	7198	0.003753	0.004925	-0.392	0.161
29B	28	345	373	0.000473	0.000244	0.959	-0.394
29C	32	440	472	0.000541	0.000311	0.801	-0.329

2AA	16	516	532	0.000271	0.000364	-0.429	0.176
2AB	0	35	35	0	0.000025	ZERO	ZERO
2AC	3	57	60	0.000051	0.00004	0.334	-0.137
2BB	1	8	9	0.000017	0.000006	1.582	-0.649
2CC	0	8	8	0	0.000006	ZERO	ZERO
333	582	28901	29483	0.00984	0.020404	-1.052	0.432
334	906	17033	17939	0.015318	0.012025	0.349	-0.143
335	28	1109	1137	0.000473	0.000783	-0.726	0.298
336	572	23899	24471	0.009671	0.016872	-0.803	0.330
337	292	11593	11885	0.004937	0.008184	-0.729	0.299
338	267	14182	14449	0.004514	0.010012	-1.149	0.472
339	1248	61411	62659	0.021101	0.043355	-1.039	0.426
33A	156	7370	7526	0.002638	0.005203	-0.980	0.402
33B	82	1252	1334	0.001386	0.000884	0.649	-0.266
33C	30	540	570	0.000507	0.000381	0.412	-0.169
344	1433	15340	16773	0.024229	0.01083	1.162	-0.477
345	4	103	107	0.000068	0.000073	-0.105	0.043
346	494	11394	11888	0.008352	0.008044	0.054	-0.022
347	228	5938	6166	0.003855	0.004192	-0.121	0.050
348	98	2979	3077	0.001657	0.002103	-0.344	0.141
349	1419	36440	37859	0.023992	0.025726	-0.101	0.041
34A	235	5815	6050	0.003973	0.004105	-0.047	0.019
34B	229	2321	2550	0.003872	0.001639	1.241	-0.509
34C	47	487	534	0.000795	0.000344	1.209	-0.496
355	0	2	2	0	0.000001	ZERO	ZERO
356	1	160	161	0.000017	0.000113	-2.740	1.125
357	0	25	25	0	0.000018	ZERO	ZERO
358	0	4	4	0	0.000003	ZERO	ZERO
359	17	979	996	0.000287	0.000691	-1.266	0.520
35A	0	44	44	0	0.000031	ZERO	ZERO
35B	0	1	1	0	0.000001	ZERO	ZERO
35C	0	1	1	0	0.000001	ZERO	ZERO
366	249	8056	8305	0.00421	0.005687	-0.434	0.178
367	173	5924	6097	0.002925	0.004182	-0.516	0.212
368	23	1058	1081	0.000389	0.000747	-0.942	0.387
369	1622	51469	53091	0.027424	0.036336	-0.406	0.167
36A	204	6980	7184	0.003449	0.004928	-0.515	0.211
36B	25	331	356	0.000423	0.000234	0.855	-0.351
36C	31	405	436	0.000524	0.000286	0.874	-0.359
377	90	2952	3042	0.001522	0.002084	-0.454	0.186
378	42	1400	1442	0.00071	0.000988	-0.477	0.196
379	480	21550	22030	0.008116	0.015214	-0.907	0.372
37A	85	2412	2497	0.001437	0.001703	-0.245	0.101
37B	22	274	296	0.000372	0.000193	0.943	-0.387
37C	11	161	172	0.000186	0.000114	0.710	-0.291
388	8	281	289	0.000135	0.000198	-0.553	0.227
389	115	10632	10747	0.001944	0.007506	-1.949	0.800
38A	40	1551	1591	0.000676	0.001095	-0.695	0.285

38B	1	92	93	0.000017	0.000065	-1.942	0.797
38C	0	91	91	0	0.000064	ZERO	ZERO
399	779	34508	35287	0.013171	0.024362	-0.887	0.364
39A	349	13213	13562	0.005901	0.009328	-0.661	0.271
39B	90	1094	1184	0.001522	0.000772	0.978	-0.401
39C	59	943	1002	0.000998	0.000666	0.583	-0.239
3AA	48	1128	1176	0.000812	0.000796	0.027	-0.011
3AB	12	155	167	0.000203	0.000109	0.891	-0.366
3AC	9	153	162	0.000152	0.000108	0.494	-0.203
3BB	8	70	78	0.000135	0.000049	1.453	-0.596
3BC	1	24	25	0.000017	0.000017	-0.003	0.001
3CC	2	32	34	0.000034	0.000023	0.582	-0.239
444	1024	6510	7534	0.017313	0.004596	1.913	-0.785
445	1	7	8	0.000017	0.000005	1.775	-0.729
446	143	2294	2437	0.002418	0.00162	0.578	-0.237
447	89	2213	2302	0.001505	0.001562	-0.054	0.022
448	9	373	382	0.000152	0.000263	-0.791	0.325
449	557	9978	10535	0.009418	0.007044	0.419	-0.172
44A	185	2513	2698	0.003128	0.001774	0.818	-0.336
44B	144	800	944	0.002435	0.000565	2.108	-0.865
44C	34	194	228	0.000575	0.000137	2.069	-0.849
456	1	16	17	0.000017	0.000011	0.582	-0.239
457	0	7	7	0	0.000005	ZERO	ZERO
458	0	1	1	0	0.000001	ZERO	ZERO
459	0	38	38	0	0.000027	ZERO	ZERO
45A	0	2	2	0	0.000001	ZERO	ZERO
45C	0	1	1	0	0.000001	ZERO	ZERO
466	52	1432	1484	0.000879	0.001011	-0.201	0.083
467	51	1636	1687	0.000862	0.001155	-0.422	0.173
468	6	128	134	0.000101	0.00009	0.167	-0.069
469	637	12763	13400	0.01077	0.00901	0.257	-0.105
46A	69	2061	2130	0.001167	0.001455	-0.319	0.131
46B	10	150	160	0.000169	0.000106	0.675	-0.277
46C	10	102	112	0.000169	0.000072	1.231	-0.505
477	53	1539	1592	0.000896	0.001087	-0.278	0.114
478	4	149	153	0.000068	0.000105	-0.637	0.261
479	202	6492	6694	0.003415	0.004583	-0.424	0.174
47A	37	1043	1080	0.000626	0.000736	-0.235	0.096
47B	14	208	222	0.000237	0.000147	0.689	-0.283
47C	3	48	51	0.000051	0.000034	0.582	-0.239
488	1	14	15	0.000017	0.00001	0.775	-0.318
489	27	1104	1131	0.000457	0.000779	-0.772	0.317
48A	23	298	321	0.000389	0.00021	0.886	-0.364
48B	0	48	48	0	0.000034	ZERO	ZERO
48C	0	9	9	0	0.000006	ZERO	ZERO
499	288	9008	9296	0.004869	0.00636	-0.385	0.158
49A	169	4932	5101	0.002857	0.003482	-0.285	0.117
49B	43	580	623	0.000727	0.000409	0.828	-0.340

49C	17	247	264	0.000287	0.000174	0.721	-0.296
4AA	37	682	719	0.000626	0.000481	0.378	-0.155
4AB	17	134	151	0.000287	0.000095	1.603	-0.658
4AC	5	53	58	0.000085	0.000037	1.176	-0.483
4BB	8	48	56	0.000135	0.000034	1.997	-0.820
4BC	2	20	22	0.000034	0.000014	1.260	-0.517
4CC	1	12	13	0.000017	0.000008	0.997	-0.409
566	0	8	8	0	0.000006	ZERO	ZERO
567	0	5	5	0	0.000004	ZERO	ZERO
568	0	1	1	0	0.000001	ZERO	ZERO
569	0	39	39	0	0.000028	ZERO	ZERO
56A	0	7	7	0	0.000005	ZERO	ZERO
577	0	1	1	0	0.000001	ZERO	ZERO
579	0	24	24	0	0.000017	ZERO	ZERO
57A	0	1	1	0	0.000001	ZERO	ZERO
589	0	5	5	0	0.000004	ZERO	ZERO
599	1	30	31	0.000017	0.000021	-0.325	0.133
59A	0	8	8	0	0.000006	ZERO	ZERO
666	11	294	305	0.000186	0.000208	-0.158	0.065
667	12	352	364	0.000203	0.000249	-0.293	0.120
668	1	37	38	0.000017	0.000026	-0.628	0.258
669	114	3675	3789	0.001927	0.002594	-0.429	0.176
66A	29	927	956	0.00049	0.000654	-0.417	0.171
66B	1	17	18	0.000017	0.000012	0.494	-0.203
66C	2	47	49	0.000034	0.000033	0.027	-0.011
677	34	681	715	0.000575	0.000481	0.258	-0.106
678	2	78	80	0.000034	0.000055	-0.704	0.289
679	128	4666	4794	0.002164	0.003294	-0.606	0.249
67A	22	604	626	0.000372	0.000426	-0.197	0.081
67B	1	26	27	0.000017	0.000018	-0.119	0.049
67C	5	36	41	0.000085	0.000025	1.734	-0.712
688	0	2	2	0	0.000001	ZERO	ZERO
689	13	529	542	0.00022	0.000373	-0.765	0.314
68A	2	67	69	0.000034	0.000047	-0.484	0.199
68B	1	3	4	0.000017	0.000002	2.997	-1.230
68C	0	6	6	0	0.000004	ZERO	ZERO
699	277	9045	9322	0.004683	0.006386	-0.447	0.183
69A	120	4025	4145	0.002029	0.002842	-0.486	0.199
69B	10	166	176	0.000169	0.000117	0.529	-0.217
69C	19	264	283	0.000321	0.000186	0.785	-0.322
6AA	13	357	370	0.00022	0.000252	-0.197	0.081
6AB	3	25	28	0.000051	0.000018	1.523	-0.625
6AC	1	46	47	0.000017	0.000032	-0.942	0.387
6BB	0	5	5	0	0.000004	ZERO	ZERO
6CC	1	7	8	0.000017	0.000005	1.775	-0.729
777	21	502	523	0.000355	0.000354	0.003	-0.001
778	4	137	141	0.000068	0.000097	-0.516	0.212
779	111	3658	3769	0.001877	0.002582	-0.461	0.189

77A	17	383	400	0.000287	0.00027	0.088	-0.036
77B	3	50	53	0.000051	0.000035	0.523	-0.215
77C	1	29	30	0.000017	0.00002	-0.276	0.113
788	0	11	11	0	0.000008	ZERO	ZERO
789	26	818	844	0.00044	0.000577	-0.394	0.162
78A	4	69	73	0.000068	0.000049	0.473	-0.194
78C	0	5	5	0	0.000004	ZERO	ZERO
799	177	6709	6886	0.002993	0.004736	-0.662	0.272
79A	64	1886	1950	0.001082	0.001331	-0.299	0.123
79B	13	103	116	0.00022	0.000073	1.596	-0.655
79C	3	91	94	0.000051	0.000064	-0.341	0.140
7AA	8	118	126	0.000135	0.000083	0.699	-0.287
7AB	2	8	10	0.000034	0.000006	2.582	-1.060
7AC	1	21	22	0.000017	0.000015	0.190	-0.078
7BB	2	4	6	0.000034	0.000003	3.582	-1.470
7BC	0	1	1	0	0.000001	ZERO	ZERO
889	2	124	126	0.000034	0.000088	-1.372	0.563
88A	1	9	10	0.000017	0.000006	1.412	-0.580
899	30	2255	2285	0.000507	0.001592	-1.650	0.677
89A	8	480	488	0.000135	0.000339	-1.325	0.544
89B	1	19	20	0.000017	0.000013	0.334	-0.137
89C	1	21	22	0.000017	0.000015	0.190	-0.078
8AA	5	51	56	0.000085	0.000036	1.231	-0.505
8AB	0	1	1	0	0.000001	ZERO	ZERO
8AC	0	11	11	0	0.000008	ZERO	ZERO
999	163	3897	4060	0.002756	0.002751	0.002	-0.001
99A	96	2874	2970	0.001623	0.002029	-0.322	0.132
99B	8	105	113	0.000135	0.000074	0.868	-0.356
99C	11	163	174	0.000186	0.000115	0.693	-0.284
9AA	31	591	622	0.000524	0.000417	0.329	-0.135
9AB	2	40	42	0.000034	0.000028	0.260	-0.107
9AC	3	89	92	0.000051	0.000063	-0.309	0.127
9BB	3	18	21	0.000051	0.000013	1.997	-0.820
9BC	0	4	4	0	0.000003	ZERO	ZERO
9CC	1	4	5	0.000017	0.000003	2.582	-1.060
AAA	10	51	61	0.000169	0.000036	2.231	-0.916
AAB	1	4	5	0.000017	0.000003	2.582	-1.060
AAC	0	6	6	0	0.000004	ZERO	ZERO
ABB	1	6	7	0.000017	0.000004	1.997	-0.820
ABC	1	1	2	0.000017	0.000001	4.582	-1.881
ACC	0	4	4	0	0.000003	ZERO	ZERO
CCC	0	4	4	0	0.000003	ZERO	ZERO

Table 9-4: The propensities and statistics for different atomic triplets in Protein-Ligand Binding Sites. ZERO indicates and division by zero result. In such a case, the triplet is given a score of 0. We also present the free energy calculations from these propensities, calculated as ΔG_{stat} = -RTln(2)×propensity.

Type	Interface Count	Surface Count	Total Count	Prob Theoretical	Prob Experimental	Log Likelihood	ΔG_{stat}
000	0	7	7	0	0.000006	ZERO	ZERO
001	219	1220	1439	0.006387	0.000998	2.679	-1.100
002	12	554	566	0.00035	0.000453	-0.372	0.153
003	64	956	1020	0.001867	0.000782	1.256	-0.515
004	12	174	186	0.00035	0.000142	1.299	-0.533
005	1	59	60	0.000029	0.000048	-0.726	0.298
006	24	885	909	0.0007	0.000724	-0.048	0.020
007	3	169	172	0.000087	0.000138	-0.659	0.271
008	0	23	23	0	0.000019	ZERO	ZERO
009	49	2464	2513	0.001429	0.002015	-0.495	0.203
00A	3	59	62	0.000087	0.000048	0.859	-0.353
00B	0	4	4	0	0.000003	ZERO	ZERO
011	381	3591	3972	0.011112	0.002936	1.920	-0.788
012	33	839	872	0.000962	0.000686	0.488	-0.200
013	431	6018	6449	0.01257	0.004921	1.353	-0.555
014	77	824	901	0.002246	0.000674	1.737	-0.713
015	0	20	20	0	0.000016	ZERO	ZERO
016	261	2694	2955	0.007612	0.002203	1.789	-0.734
017	60	766	826	0.00175	0.000626	1.482	-0.608
018	2	89	91	0.000058	0.000073	-0.319	0.131
019	145	3044	3189	0.004229	0.002489	0.765	-0.314
01A	234	3232	3466	0.006825	0.002643	1.369	-0.562
01B	0	16	16	0	0.000013	ZERO	ZERO
01C	3	12	15	0.000087	0.00001	3.157	-1.296
022	10	474	484	0.000292	0.000388	-0.410	0.168
023	136	9375	9511	0.003966	0.007666	-0.951	0.390
024	70	3387	3457	0.002042	0.002769	-0.440	0.181
025	18	1174	1192	0.000525	0.00096	-0.871	0.357
026	315	15471	15786	0.009187	0.01265	-0.461	0.189
027	19	1386	1405	0.000554	0.001133	-1.032	0.424
028	3	187	190	0.000087	0.000153	-0.805	0.331
029	452	36134	36586	0.013182	0.029545	-1.164	0.478
02A	28	1632	1660	0.000817	0.001334	-0.708	0.291
02B	0	44	44	0	0.000036	ZERO	ZERO
02C	1	41	42	0.000029	0.000034	-0.201	0.082
033	129	8673	8802	0.003762	0.007092	-0.915	0.375
034	137	3894	4031	0.003996	0.003184	0.328	-0.134
035	8	621	629	0.000233	0.000508	-1.122	0.460
036	405	20190	20595	0.011812	0.016508	-0.483	0.198
037	212	14046	14258	0.006183	0.011485	-0.893	0.367
038	15	723	738	0.000437	0.000591	-0.434	0.178
039	796	65791	66587	0.023215	0.053794	-1.212	0.498

03A	58	2210	2268	0.001692	0.001807	-0.095	0.039
03B	2	80	82	0.000058	0.000065	-0.165	0.068
03C	8	62	70	0.000233	0.000051	2.202	-0.904
044	27	664	691	0.000787	0.000543	0.536	-0.220
045	2	55	57	0.000058	0.000045	0.375	-0.154
046	70	2411	2481	0.002042	0.001971	0.050	-0.021
047	17	818	835	0.000496	0.000669	-0.432	0.177
048	0	40	40	0	0.000033	ZERO	ZERO
049	169	8577	8746	0.004929	0.007013	-0.509	0.209
04A	18	564	582	0.000525	0.000461	0.187	-0.077
04B	0	27	27	0	0.000022	ZERO	ZERO
04C	0	13	13	0	0.000011	ZERO	ZERO
055	0	2	2	0	0.000002	ZERO	ZERO
056	4	131	135	0.000117	0.000107	0.123	-0.051
057	0	20	20	0	0.000016	ZERO	ZERO
058	0	4	4	0	0.000003	ZERO	ZERO
059	5	448	453	0.000146	0.000366	-1.329	0.545
05A	1	25	26	0.000029	0.00002	0.513	-0.210
066	33	957	990	0.000962	0.000782	0.299	-0.123
067	86	5602	5688	0.002508	0.004581	-0.869	0.357
068	4	125	129	0.000117	0.000102	0.191	-0.078
069	247	13345	13592	0.007204	0.010912	-0.599	0.246
06A	23	812	835	0.000671	0.000664	0.015	-0.006
06B	2	24	26	0.000058	0.00002	1.572	-0.645
06C	0	17	17	0	0.000014	ZERO	ZERO
077	80	5461	5541	0.002333	0.004465	-0.936	0.384
078	1	125	126	0.000029	0.000102	-1.809	0.743
079	214	12717	12931	0.006241	0.010398	-0.736	0.302
07A	18	464	482	0.000525	0.000379	0.469	-0.192
07B	0	16	16	0	0.000013	ZERO	ZERO
07C	0	8	8	0	0.000007	ZERO	ZERO
088	1	4	5	0.000029	0.000003	3.157	-1.296
089	12	1480	1492	0.00035	0.00121	-1.790	0.735
08A	0	33	33	0	0.000027	ZERO	ZERO
099	327	27734	28061	0.009537	0.022677	-1.250	0.513
09A	72	3164	3236	0.0021	0.002587	-0.301	0.124
09B	4	113	117	0.000117	0.000092	0.336	-0.138
09C	3	65	68	0.000087	0.000053	0.719	-0.295
0AA	7	99	106	0.000204	0.000081	1.335	-0.548
0AB	0	2	2	0	0.000002	ZERO	ZERO
0AC	0	2	2	0	0.000002	ZERO	ZERO
0BB	0	4	4	0	0.000003	ZERO	ZERO
111	951	7172	8123	0.027736	0.005864	2.242	-0.920
112	78	1530	1608	0.002275	0.001251	0.863	-0.354
113	442	7701	8143	0.012891	0.006297	1.034	-0.424
114	638	5329	5967	0.018607	0.004357	2.094	-0.860
115	0	6	6	0	0.000005	ZERO	ZERO
116	384	4949	5333	0.011199	0.004047	1.469	-0.603

117	162	2084	2246	0.004725	0.001704	1.471	-0.604
118	18	428	446	0.000525	0.00035	0.585	-0.240
119	294	6749	7043	0.008574	0.005518	0.636	-0.261
11A	176	1739	1915	0.005133	0.001422	1.852	-0.760
11B	18	149	167	0.000525	0.000122	2.107	-0.865
11C	38	117	155	0.001108	0.000096	3.534	-1.451
122	15	297	312	0.000437	0.000243	0.849	-0.349
123	207	5906	6113	0.006037	0.004829	0.322	-0.132
124	192	3026	3218	0.0056	0.002474	1.178	-0.484
125	2	20	22	0.000058	0.000016	1.835	-0.753
126	142	3735	3877	0.004141	0.003054	0.439	-0.180
127	23	509	532	0.000671	0.000416	0.689	-0.283
128	3	107	110	0.000087	0.000087	0.000	0.000
129	185	7117	7302	0.005395	0.005819	-0.109	0.045
12A	47	948	995	0.001371	0.000775	0.822	-0.338
12B	5	34	39	0.000146	0.000028	2.391	-0.981
12C	8	39	47	0.000233	0.000032	2.871	-1.179
133	253	7321	7574	0.007379	0.005986	0.302	-0.124
134	419	6123	6542	0.01222	0.005007	1.287	-0.528
135	1	36	37	0.000029	0.000029	-0.013	0.005
136	383	8290	8673	0.01117	0.006778	0.721	-0.296
137	173	2803	2976	0.005045	0.002292	1.138	-0.467
138	41	1273	1314	0.001196	0.001041	0.200	-0.082
139	461	14501	14962	0.013445	0.011857	0.181	-0.074
13A	222	4158	4380	0.006475	0.0034	0.929	-0.381
13B	32	307	339	0.000933	0.000251	1.894	-0.778
13C	38	149	187	0.001108	0.000122	3.185	-1.307
144	464	3635	4099	0.013532	0.002972	2.187	-0.898
145	0	1	1	0	0.000001	ZERO	ZERO
146	191	3028	3219	0.00557	0.002476	1.170	-0.480
147	82	986	1068	0.002392	0.000806	1.569	-0.644
148	9	131	140	0.000262	0.000107	1.293	-0.531
149	202	4578	4780	0.005891	0.003743	0.654	-0.269
14A	188	2625	2813	0.005483	0.002146	1.353	-0.555
14B	43	310	353	0.001254	0.000253	2.307	-0.947
14C	20	80	100	0.000583	0.000065	3.157	-1.296
156	0	2	2	0	0.000002	ZERO	ZERO
157	0	1	1	0	0.000001	ZERO	ZERO
159	0	15	15	0	0.000012	ZERO	ZERO
15A	1	1	2	0.000029	0.000001	5.157	-2.117
166	97	1048	1145	0.002829	0.000857	1.723	-0.707
167	149	1710	1859	0.004346	0.001398	1.636	-0.672
168	22	270	292	0.000642	0.000221	1.539	-0.632
169	327	6912	7239	0.009537	0.005652	0.755	-0.310
16A	106	1188	1294	0.003091	0.000971	1.670	-0.686
16B	4	79	83	0.000117	0.000065	0.853	-0.350
16C	6	34	40	0.000175	0.000028	2.654	-1.089
177	79	754	833	0.002304	0.000617	1.902	-0.781

178	1	47	48	0.000029	0.000038	-0.398	0.163
179	183	2619	2802	0.005337	0.002141	1.317	-0.541
17A	89	897	986	0.002596	0.000733	1.823	-0.748
17B	7	25	32	0.000204	0.00002	3.320	-1.363
17C	2	17	19	0.000058	0.000014	2.069	-0.849
188	0	5	5	0	0.000004	ZERO	ZERO
189	26	533	559	0.000758	0.000436	0.799	-0.328
18A	13	267	280	0.000379	0.000218	0.796	-0.327
18B	1	6	7	0.000029	0.000005	2.572	-1.056
18C	0	1	1	0	0.000001	ZERO	ZERO
199	152	2999	3151	0.004433	0.002452	0.854	-0.351
19A	142	2555	2697	0.004141	0.002089	0.987	-0.405
19B	5	83	88	0.000146	0.000068	1.103	-0.453
19C	8	49	57	0.000233	0.00004	2.542	-1.043
1AA	36	322	358	0.00105	0.000263	1.996	-0.819
1AB	5	51	56	0.000146	0.000042	1.806	-0.741
1AC	6	32	38	0.000175	0.000026	2.742	-1.125
1BB	1	8	9	0.000029	0.000007	2.157	-0.885
1BC	0	5	5	0	0.000004	ZERO	ZERO
222	0	66	66	0	0.000054	ZERO	ZERO
223	50	1561	1611	0.001458	0.001276	0.192	-0.079
224	110	4673	4783	0.003208	0.003821	-0.252	0.104
225	0	26	26	0	0.000021	ZERO	ZERO
226	69	2305	2374	0.002012	0.001885	0.095	-0.039
227	3	194	197	0.000087	0.000159	-0.858	0.352
228	1	46	47	0.000029	0.000038	-0.367	0.151
229	114	5041	5155	0.003325	0.004122	-0.310	0.127
22A	32	962	994	0.000933	0.000787	0.247	-0.101
22B	0	21	21	0	0.000017	ZERO	ZERO
22C	3	22	25	0.000087	0.000018	2.282	-0.937
233	250	22535	22785	0.007291	0.018426	-1.338	0.549
234	486	18194	18680	0.014174	0.014876	-0.070	0.029
235	21	1877	1898	0.000612	0.001535	-1.325	0.544
236	598	30034	30632	0.017441	0.024557	-0.494	0.203
237	81	4783	4864	0.002362	0.003911	-0.727	0.299
238	32	1408	1440	0.000933	0.001151	-0.303	0.124
239	666	62708	63374	0.019424	0.051274	-1.400	0.575
23A	130	5459	5589	0.003791	0.004464	-0.235	0.097
23B	27	443	470	0.000787	0.000362	1.120	-0.460
23C	14	287	301	0.000408	0.000235	0.799	-0.328
244	529	11666	12195	0.015428	0.009539	0.694	-0.285
245	0	91	91	0	0.000074	ZERO	ZERO
246	373	14495	14868	0.010878	0.011852	-0.124	0.051
247	68	2442	2510	0.001983	0.001997	-0.010	0.004
248	8	441	449	0.000233	0.000361	-0.628	0.258
249	539	29778	30317	0.01572	0.024348	-0.631	0.259
24A	154	7134	7288	0.004491	0.005833	-0.377	0.155
24B	18	300	318	0.000525	0.000245	1.098	-0.451

24C	12	154	166	0.00035	0.000126	1.475	-0.605
256	1	40	41	0.000029	0.000033	-0.165	0.068
257	0	24	24	0	0.00002	ZERO	ZERO
258	0	4	4	0	0.000003	ZERO	ZERO
259	1	416	417	0.000029	0.00034	-3.544	1.455
25A	1	35	36	0.000029	0.000029	0.027	-0.011
25C	0	1	1	0	0.000001	ZERO	ZERO
266	171	3671	3842	0.004987	0.003002	0.732	-0.301
267	66	2452	2518	0.001925	0.002005	-0.059	0.024
268	39	446	485	0.001137	0.000365	1.641	-0.674
269	536	25009	25545	0.015632	0.020449	-0.387	0.159
26A	163	3934	4097	0.004754	0.003217	0.564	-0.231
26B	7	176	183	0.000204	0.000144	0.505	-0.207
26C	15	110	125	0.000437	0.00009	2.282	-0.937
277	6	355	361	0.000175	0.00029	-0.730	0.300
278	4	54	58	0.000117	0.000044	1.402	-0.575
279	104	6205	6309	0.003033	0.005074	-0.742	0.305
27A	19	550	569	0.000554	0.00045	0.301	-0.124
27B	0	18	18	0	0.000015	ZERO	ZERO
27C	0	19	19	0	0.000016	ZERO	ZERO
288	0	7	7	0	0.000006	ZERO	ZERO
289	22	1830	1852	0.000642	0.001496	-1.222	0.501
28A	7	130	137	0.000204	0.000106	0.942	-0.386
28B	0	10	10	0	0.000008	ZERO	ZERO
28C	0	3	3	0	0.000002	ZERO	ZERO
299	242	19679	19921	0.007058	0.016091	-1.189	0.488
29A	128	5523	5651	0.003733	0.004516	-0.275	0.113
29B	12	329	341	0.00035	0.000269	0.380	-0.156
29C	11	189	200	0.000321	0.000155	1.054	-0.433
2AA	18	457	475	0.000525	0.000374	0.490	-0.201
2AB	2	44	46	0.000058	0.000036	0.697	-0.286
2AC	2	15	17	0.000058	0.000012	2.250	-0.923
2BB	1	14	15	0.000029	0.000011	1.349	-0.554
2BC	0	2	2	0	0.000002	ZERO	ZERO
2CC	0	1	1	0	0.000001	ZERO	ZERO
333	311	25122	25433	0.00907	0.020541	-1.179	0.484
334	451	14324	14775	0.013153	0.011712	0.167	-0.069
335	6	989	995	0.000175	0.000809	-2.208	0.906
336	478	20785	21263	0.013941	0.016995	-0.286	0.117
337	228	9986	10214	0.00665	0.008165	-0.296	0.122
338	158	12192	12350	0.004608	0.009969	-1.113	0.457
339	663	54617	55280	0.019336	0.044658	-1.208	0.496
33A	146	5987	6133	0.004258	0.004895	-0.201	0.083
33B	50	937	987	0.001458	0.000766	0.929	-0.381
33C	18	272	290	0.000525	0.000222	1.239	-0.509
344	695	12512	13207	0.020269	0.010231	0.986	-0.405
345	2	84	86	0.000058	0.000069	-0.236	0.097
346	412	10279	10691	0.012016	0.008405	0.516	-0.212

347	155	5132	5287	0.004521	0.004196	0.107	-0.044
348	54	2295	2349	0.001575	0.001877	-0.253	0.104
349	706	32374	33080	0.02059	0.026471	-0.362	0.149
34A	221	4941	5162	0.006445	0.00404	0.674	-0.277
34B	111	1745	1856	0.003237	0.001427	1.182	-0.485
34C	29	247	276	0.000846	0.000202	2.066	-0.848
355	0	2	2	0	0.000002	ZERO	ZERO
356	3	151	154	0.000087	0.000123	-0.497	0.204
357	0	14	14	0	0.000011	ZERO	ZERO
358	0	4	4	0	0.000003	ZERO	ZERO
359	5	817	822	0.000146	0.000668	-2.196	0.901
35A	1	36	37	0.000029	0.000029	-0.013	0.005
366	466	7702	8168	0.013591	0.006298	1.110	-0.456
367	180	5476	5656	0.00525	0.004477	0.230	-0.094
368	85	1047	1132	0.002479	0.000856	1.534	-0.630
369	927	46342	47269	0.027036	0.037892	-0.487	0.200
36A	326	6025	6351	0.009508	0.004926	0.949	-0.389
36B	17	330	347	0.000496	0.00027	0.878	-0.360
36C	26	240	266	0.000758	0.000196	1.950	-0.800
377	94	2698	2792	0.002741	0.002206	0.314	-0.129
378	38	1072	1110	0.001108	0.000877	0.338	-0.139
379	373	20048	20421	0.010878	0.016392	-0.592	0.243
37A	80	2076	2156	0.002333	0.001697	0.459	-0.188
37B	15	180	195	0.000437	0.000147	1.572	-0.645
37C	8	88	96	0.000233	0.000072	1.697	-0.697
388	6	206	212	0.000175	0.000168	0.055	-0.023
389	152	10486	10638	0.004433	0.008574	-0.952	0.391
38A	49	1289	1338	0.001429	0.001054	0.439	-0.180
38B	0	105	105	0	0.000086	ZERO	ZERO
38C	0	34	34	0	0.000028	ZERO	ZERO
399	486	30997	31483	0.014174	0.025345	-0.838	0.344
39A	277	10983	11260	0.008079	0.00898	-0.153	0.063
39B	24	873	897	0.0007	0.000714	-0.028	0.012
39C	33	511	544	0.000962	0.000418	1.204	-0.494
3AA	40	792	832	0.001167	0.000648	0.849	-0.349
3AB	10	138	148	0.000292	0.000113	1.370	-0.562
3AC	13	69	82	0.000379	0.000056	2.749	-1.128
3BB	6	83	89	0.000175	0.000068	1.367	-0.561
3BC	3	21	24	0.000087	0.000017	2.349	-0.964
3CC	0	13	13	0	0.000011	ZERO	ZERO
444	604	5644	6248	0.017615	0.004615	1.932	-0.793
445	0	1	1	0	0.000001	ZERO	ZERO
446	124	2496	2620	0.003616	0.002041	0.825	-0.339
447	95	1940	2035	0.002771	0.001586	0.805	-0.330
448	4	257	261	0.000117	0.00021	-0.849	0.349
449	357	9524	9881	0.010412	0.007787	0.419	-0.172
44A	141	2300	2441	0.004112	0.001881	1.129	-0.463
44B	87	616	703	0.002537	0.000504	2.333	-0.958

44C	18	113	131	0.000525	0.000092	2.506	-1.029
456	0	13	13	0	0.000011	ZERO	ZERO
459	0	36	36	0	0.000029	ZERO	ZERO
466	65	1446	1511	0.001896	0.001182	0.681	-0.280
467	77	1626	1703	0.002246	0.00133	0.756	-0.310
468	1	113	114	0.000029	0.000092	-1.664	0.683
469	337	12456	12793	0.009829	0.010185	-0.051	0.021
46A	98	1985	2083	0.002858	0.001623	0.816	-0.335
46B	7	138	145	0.000204	0.000113	0.855	-0.351
46C	9	68	77	0.000262	0.000056	2.239	-0.919
477	62	1307	1369	0.001808	0.001069	0.759	-0.311
478	5	92	97	0.000146	0.000075	0.955	-0.392
479	174	6193	6367	0.005075	0.005064	0.003	-0.001
47A	40	838	878	0.001167	0.000685	0.768	-0.315
47B	9	154	163	0.000262	0.000126	1.060	-0.435
47C	1	29	30	0.000029	0.000024	0.299	-0.123
488	0	2	2	0	0.000002	ZERO	ZERO
489	28	1186	1214	0.000817	0.00097	-0.248	0.102
48A	5	221	226	0.000146	0.000181	-0.309	0.127
48B	4	43	47	0.000117	0.000035	1.730	-0.710
48C	0	8	8	0	0.000007	ZERO	ZERO
499	231	8429	8660	0.006737	0.006892	-0.033	0.013
49A	133	4208	4341	0.003879	0.003441	0.173	-0.071
49B	25	508	533	0.000729	0.000415	0.812	-0.333
49C	14	174	188	0.000408	0.000142	1.521	-0.624
4AA	23	511	534	0.000671	0.000418	0.683	-0.280
4AB	14	110	124	0.000408	0.00009	2.183	-0.896
4AC	6	38	44	0.000175	0.000031	2.494	-1.024
4BB	3	46	49	0.000087	0.000038	1.218	-0.500
4BC	1	9	10	0.000029	0.000007	1.987	-0.815
4CC	1	5	6	0.000029	0.000004	2.835	-1.164
566	0	8	8	0	0.000007	ZERO	ZERO
567	0	3	3	0	0.000002	ZERO	ZERO
568	0	2	2	0	0.000002	ZERO	ZERO
569	0	27	27	0	0.000022	ZERO	ZERO
56A	0	5	5	0	0.000004	ZERO	ZERO
579	0	12	12	0	0.00001	ZERO	ZERO
599	0	19	19	0	0.000016	ZERO	ZERO
59A	0	14	14	0	0.000011	ZERO	ZERO
59C	0	1	1	0	0.000001	ZERO	ZERO
666	40	310	350	0.001167	0.000253	2.202	-0.904
667	40	367	407	0.001167	0.0003	1.959	-0.804
668	56	115	171	0.001633	0.000094	4.118	-1.691
669	146	3820	3966	0.004258	0.003123	0.447	-0.184
66A	85	896	981	0.002479	0.000733	1.759	-0.722
66B	2	21	23	0.000058	0.000017	1.764	-0.724
66C	6	39	45	0.000175	0.000032	2.456	-1.008
677	50	617	667	0.001458	0.000504	1.531	-0.629

678	5	69	74	0.000146	0.000056	1.370	-0.562
679	134	5060	5194	0.003908	0.004137	-0.082	0.034
67A	39	531	570	0.001137	0.000434	1.389	-0.570
67B	3	32	35	0.000087	0.000026	1.742	-0.715
67C	3	25	28	0.000087	0.00002	2.098	-0.861
688	0	1	1	0	0.000001	ZERO	ZERO
689	12	540	552	0.00035	0.000442	-0.335	0.138
68A	4	93	97	0.000117	0.000076	0.617	-0.253
68B	0	3	3	0	0.000002	ZERO	ZERO
699	235	9065	9300	0.006854	0.007412	-0.113	0.046
69A	118	3487	3605	0.003441	0.002851	0.271	-0.111
69B	5	143	148	0.000146	0.000117	0.319	-0.131
69C	13	155	168	0.000379	0.000127	1.581	-0.649
6AA	19	258	277	0.000554	0.000211	1.393	-0.572
6AB	3	32	35	0.000087	0.000026	1.742	-0.715
6AC	8	36	44	0.000233	0.000029	2.987	-1.226
6BB	0	7	7	0	0.000006	ZERO	ZERO
6BC	0	1	1	0	0.000001	ZERO	ZERO
6CC	1	2	3	0.000029	0.000002	4.157	-1.706
777	43	440	483	0.001254	0.00036	1.801	-0.739
778	6	77	83	0.000175	0.000063	1.475	-0.605
779	106	3757	3863	0.003091	0.003072	0.009	-0.004
77A	46	365	411	0.001342	0.000298	2.168	-0.890
77B	4	21	25	0.000117	0.000017	2.764	-1.135
77C	0	14	14	0	0.000011	ZERO	ZERO
788	0	2	2	0	0.000002	ZERO	ZERO
789	18	775	793	0.000525	0.000634	-0.272	0.111
78A	3	52	55	0.000087	0.000043	1.041	-0.427
799	209	6507	6716	0.006095	0.00532	0.196	-0.081
79A	63	1610	1673	0.001837	0.001316	0.481	-0.197
79B	5	78	83	0.000146	0.000064	1.193	-0.490
79C	2	59	61	0.000058	0.000048	0.274	-0.112
7AA	6	91	97	0.000175	0.000074	1.234	-0.506
7AB	0	5	5	0	0.000004	ZERO	ZERO
7AC	1	5	6	0.000029	0.000004	2.835	-1.164
7BB	0	4	4	0	0.000003	ZERO	ZERO
888	0	2	2	0	0.000002	ZERO	ZERO
889	8	97	105	0.000233	0.000079	1.557	-0.639
88A	0	2	2	0	0.000002	ZERO	ZERO
899	52	2381	2433	0.001517	0.001947	-0.360	0.148
89A	12	508	520	0.00035	0.000415	-0.247	0.101
89B	0	28	28	0	0.000023	ZERO	ZERO
89C	0	10	10	0	0.000008	ZERO	ZERO
8AA	1	20	21	0.000029	0.000016	0.835	-0.343
8AB	0	7	7	0	0.000006	ZERO	ZERO
8CC	0	1	1	0	0.000001	ZERO	ZERO
999	146	3456	3602	0.004258	0.002826	0.592	-0.243
99A	72	2478	2550	0.0021	0.002026	0.052	-0.021

99B	8	88	96	0.000233	0.000072	1.697	-0.697
99C	7	60	67	0.000204	0.000049	2.057	-0.844
9AA	25	452	477	0.000729	0.00037	0.980	-0.402
9AB	1	35	36	0.000029	0.000029	0.027	-0.011
9AC	6	41	47	0.000175	0.000034	2.384	-0.979
9BB	1	27	28	0.000029	0.000022	0.402	-0.165
9BC	1	2	3	0.000029	0.000002	4.157	-1.706
9CC	0	7	7	0	0.000006	ZERO	ZERO
AAA	1	14	15	0.000029	0.000011	1.349	-0.554
AAB	0	5	5	0	0.000004	ZERO	ZERO
AAC	0	5	5	0	0.000004	ZERO	ZERO
ABB	0	3	3	0	0.000002	ZERO	ZERO
ABC	1	4	5	0.000029	0.000003	3.157	-1.296
CCC	0	1	1	0	0.000001	ZERO	ZERO

Table 9-5: The propensities and statistics for different atomic groups in Protein-Protein Binding Sites. ZERO indicates and division by zero. In such a case, the triplet is given a score of 0

Type	Interface Count	Surface Count	Total Count	Prob Theoretical	Prob Experimental	Log Likelihood
0	14118	55026	69144	0.092365	0.08466	0.125668
1	17797	34843	52640	0.116434	0.053608	1.119011
2	18493	74934	93427	0.120988	0.115289	0.069604
3	32487	163490	195977	0.212542	0.251537	-0.243022
4	21586	63314	84900	0.141223	0.097411	0.535817
5	244	1245	1489	0.001596	0.001915	-0.262947
6	12842	58400	71242	0.084017	0.089851	-0.096853
7	5055	27939	32994	0.033072	0.042985	-0.378252
8	610	8202	8812	0.003991	0.012619	-1.660849
9	24326	139890	164216	0.159149	0.215227	-0.435476
A	4100	20029	24129	0.026824	0.030816	-0.200149
B	871	2106	2977	0.005698	0.00324	0.814485
C	321	547	868	0.0021	0.000842	1.319278

Table 9-6: The propensities and statistics for different atomic groups in Protein-Peptide Binding Sites. ZERO indicates and division by zero. In such a case, the triplet is given a score of 0

Type	Interface Count	Surface Count	Total Count	Prob Theoretical	Prob Experimental	Log Likelihood
0	4027	80924	84951	0.106049	0.124419	-0.23048
1	4686	38302	42988	0.123403	0.058889	1.067319
2	4182	85733	89915	0.110131	0.131813	-0.259275
3	7391	143191	150582	0.194638	0.220154	-0.177717
4	4711	50931	55642	0.124062	0.078306	0.663871
5	49	1949	1998	0.00129	0.002997	-1.215498
6	3354	68206	71560	0.088326	0.104866	-0.247634
7	1510	24652	26162	0.039765	0.037902	0.069226
8	352	8189	8541	0.00927	0.01259	-0.44173
9	6083	126481	132564	0.160193	0.194463	-0.279684
A	1296	19008	20304	0.03413	0.029225	0.223841
B	239	1865	2104	0.006294	0.002867	1.134217
C	93	982	1075	0.002449	0.00151	0.69789

Table 9-7: The propensities and statistics for different atomic groups in Protein-Ligand Binding Sites. ZERO indicates and division by zero result which is not calculatable. In such a case, the triplet is given a score of 0

Type	Interface Count	Surface Count	Total Count	Prob Theoretical	Prob Experimental	Log Likelihood
0	1773	68498	70271	0.086665	0.121607	-0.4887
1	2867	31897	34764	0.140141	0.056628	1.30729
2	1807	72955	74762	0.088327	0.12952	-0.552241
3	3833	124494	128327	0.187359	0.221019	-0.238361
4	2584	44633	47217	0.126308	0.079239	0.672664
5	22	1737	1759	0.001075	0.003084	-1.519853
6	2325	60661	62986	0.113647	0.107694	0.077631
7	967	21820	22787	0.047268	0.038738	0.287106
8	216	7321	7537	0.010558	0.012997	-0.29984
9	3018	111685	114703	0.147522	0.198279	-0.426602
A	851	15539	16390	0.041597	0.027587	0.592507
B	127	1516	1643	0.006208	0.002691	1.205728
C	68	517	585	0.003324	0.000918	1.85654

10 Appendix B: The C5/C7 complex docking with Hex and STP

Table 10-1: Docking Experiment 1/3: Hex docking orientations and STP Scores - The C5/C7 complex

Dock Rank	Energy	T-Score	STP Rank	Dock Rank	Energy	T-Score	STP Rank
0001	-5.03E+02	-0.523532	449	0251	-2.13E+02	-0.243663	227
0002	-5.02E+02	-0.430512	373	0252	-2.12E+02	-0.062308	97
0003	-4.83E+02	0.009458	61	0253	-2.12E+02	-0.387508	340
0004	-4.77E+02	-0.16933	163	0254	-2.12E+02	-0.507656	440
0005	-4.57E+02	-0.273236	251	0255	-2.11E+02	-0.105118	126
0006	-4.56E+02	-0.566823	462	0256	-2.10E+02	-0.389233	342
0007	-4.52E+02	-0.06718	98	0257	-2.09E+02	-0.3869	339
0008	-4.50E+02	0.112021	20	0258	-2.08E+02	-0.279266	257
0009	-4.46E+02	-0.599149	469	0259	-2.08E+02	-0.425419	370
0010	-4.39E+02	-0.171933	164	0260	-2.08E+02	-0.611392	477
0011	-4.36E+02	-0.282511	259	0261	-2.08E+02	-0.496613	434
0012	-4.24E+02	-0.042761	87	0262	-2.08E+02	-0.542941	456
0013	-4.19E+02	-0.221377	208	0263	-2.07E+02	-0.195647	188
0014	-4.19E+02	-0.001205	68	0264	-2.07E+02	-0.507285	439
0015	-4.17E+02	-0.466396	412	0265	-2.05E+02	-0.424212	367
0016	-4.14E+02	-0.334434	304	0266	-2.05E+02	-0.495699	433
0017	-4.10E+02	-0.499116	436	0267	-2.04E+02	-0.599645	470
0018	-4.04E+02	-0.421603	365	0268	-2.04E+02	-0.437271	380
0019	-4.02E+02	-0.185757	174	0269	-2.04E+02	0.099506	25
0020	-4.01E+02	-0.358034	317	0270	-2.03E+02	-0.174138	167
0021	-4.00E+02	-0.133862	143	0271	-2.03E+02	-0.075016	103
0022	-3.99E+02	-0.450014	393	0272	-2.03E+02	-0.362419	320
0023	-3.91E+02	-0.27811	255	0273	-2.03E+02	-0.178149	168
0024	-3.91E+02	-0.321193	292	0274	-2.03E+02	-0.143461	149
0025	-3.90E+02	-0.570597	464	0275	-2.02E+02	0.026898	54
0026	-3.88E+02	-0.018964	74	0276	-2.02E+02	0.023583	56
0027	-3.87E+02	-0.478564	419	0277	-2.02E+02	-0.769574	500
0028	-3.85E+02	-0.601326	471	0278	-2.02E+02	-0.208422	198
0029	-3.85E+02	-0.487037	427	0279	-2.02E+02	-0.457464	404
0030	-3.83E+02	-0.328887	299	0280	-2.01E+02	-0.445494	391
0031	-3.81E+02	-0.363649	323	0281	-2.01E+02	-0.267563	249
0032	-3.81E+02	-0.224	212	0282	-2.00E+02	-0.22229	210
0033	-3.80E+02	-0.117946	135	0283	-2.00E+02	-0.095229	120
0034	-3.76E+02	-0.470685	414	0284	-2.00E+02	-0.241358	225
0035	-3.76E+02	0.215978	4	0285	-2.00E+02	-0.391344	344
0036	-3.74E+02	-0.653767	488	0286	-2.00E+02	-0.164288	160
0037	-3.71E+02	0.135823	12	0287	-1.99E+02	-0.071529	100

0038	-3.66E+02	-0.044305	91	0288	-1.99E+02	-0.299289	268
0039	-3.65E+02	-0.290992	263	0289	-1.98E+02	-0.439924	384
0040	-3.65E+02	-0.432284	376	0290	-1.97E+02	-0.291613	264
0041	-3.64E+02	-0.211555	201	0291	-1.96E+02	-0.245517	230
0042	-3.62E+02	-0.52794	450	0292	-1.96E+02	-0.203996	194
0043	-3.61E+02	-0.396593	348	0293	-1.96E+02	0.200308	5
0044	-3.60E+02	-0.42813	371	0294	-1.96E+02	-0.201491	190
0045	-3.60E+02	-0.553282	459	0295	-1.95E+02	-0.257963	242
0046	-3.59E+02	-0.023047	77	0296	-1.95E+02	-0.227137	215
0047	-3.59E+02	-0.333151	302	0297	-1.95E+02	-0.338594	306
0048	-3.59E+02	-0.553637	460	0298	-1.94E+02	-0.346494	309
0049	-3.58E+02	-0.462201	406	0299	-1.94E+02	-0.186974	180
0050	-3.56E+02	0.027352	53	0300	-1.93E+02	-0.263556	246
0051	-3.49E+02	-0.013315	72	0301	-1.93E+02	-0.663403	491
0052	-3.48E+02	-0.4377	381	0302	-1.91E+02	-0.30111	270
0053	-3.48E+02	-0.183046	172	0303	-1.91E+02	-0.236884	221
0054	-3.47E+02	-0.522058	447	0304	-1.91E+02	-0.334061	303
0055	-3.46E+02	-0.194725	186	0305	-1.90E+02	-0.21437	202
0056	-3.40E+02	-0.108044	129	0306	-1.90E+02	-0.208326	197
0057	-3.38E+02	-0.441933	387	0307	-1.90E+02	-0.084844	110
0058	-3.37E+02	-0.706253	496	0308	-1.90E+02	-0.464405	411
0059	-3.36E+02	-0.250228	234	0309	-1.89E+02	-0.315081	283
0060	-3.35E+02	0.0692	32	0310	-1.89E+02	0.1303	14
0061	-3.35E+02	0.002661	65	0311	-1.88E+02	-0.41054	357
0062	-3.33E+02	-0.601986	472	0312	-1.88E+02	-0.025214	80
0063	-3.33E+02	-0.190409	182	0313	-1.87E+02	-0.404457	354
0064	-3.32E+02	-0.195493	187	0314	-1.86E+02	-0.456788	403
0065	-3.32E+02	-0.130902	139	0315	-1.86E+02	-0.510488	442
0066	-3.31E+02	-0.201559	191	0316	-1.86E+02	-0.090026	116
0067	-3.31E+02	-0.424597	368	0317	-1.86E+02	-0.436856	379
0068	-3.29E+02	-0.319402	288	0318	-1.85E+02	-0.014283	73
0069	-3.27E+02	-0.091589	118	0319	-1.85E+02	-0.467187	413
0070	-3.27E+02	-0.259283	243	0320	-1.85E+02	-0.455184	400
0071	-3.26E+02	-0.633207	481	0321	-1.84E+02	-0.636845	483
0072	-3.25E+02	-0.409281	356	0322	-1.83E+02	-0.311231	277
0073	-3.25E+02	-0.347563	310	0323	-1.81E+02	-0.498805	435
0074	-3.23E+02	-0.113094	132	0324	-1.81E+02	-0.191805	183
0075	-3.20E+02	-0.225254	214	0325	-1.80E+02	-0.237871	222
0076	-3.19E+02	-0.444053	390	0326	-1.78E+02	-0.484774	424
0077	-3.19E+02	-0.462526	407	0327	-1.77E+02	0.13651	11
0078	-3.18E+02	-0.149569	151	0328	-1.75E+02	-0.185822	175
0079	-3.18E+02	-0.074309	102	0329	-1.74E+02	-0.185041	173
0080	-3.17E+02	-0.077047	106	0330	-1.73E+02	-0.267286	248
0081	-3.16E+02	-0.419252	362	0331	-1.73E+02	0.029433	52
0082	-3.14E+02	-0.367962	327	0332	-1.72E+02	-0.077722	108
0083	-3.14E+02	-0.290516	262	0333	-1.71E+02	-0.061819	96
0084	-3.13E+02	-0.325818	298	0334	-1.70E+02	0.037525	46
0085	-3.11E+02	-0.394879	347	0335	-1.69E+02	-0.241286	224

0086	-3.09E+02	-0.18798	181	0336	-1.69E+02	-0.474176	417
0087	-3.08E+02	0.127471	15	0337	-1.68E+02	0.036	48
0088	-3.08E+02	-0.247431	233	0338	-1.68E+02	-0.705911	495
0089	-3.07E+02	-0.0248	78	0339	-1.67E+02	-0.387508	340
0090	-3.07E+02	-0.419615	363	0340	-1.67E+02	-0.376668	330
0091	-3.07E+02	-0.454123	398	0341	-1.67E+02	-0.341373	308
0092	-3.04E+02	-0.16198	157	0342	-1.67E+02	-0.607551	475
0093	-3.03E+02	-0.453504	397	0343	-1.66E+02	-0.509811	441
0094	-3.03E+02	-0.649187	487	0344	-1.66E+02	-0.479508	420
0095	-3.02E+02	-0.203195	193	0345	-1.65E+02	-0.384765	335
0096	-3.01E+02	-0.193274	184	0346	-1.64E+02	-0.450489	394
0097	-3.00E+02	0.040655	44	0347	-1.64E+02	-0.312248	278
0098	-2.99E+02	0.115836	17	0348	-1.64E+02	0.113825	19
0099	-2.97E+02	0.058657	38	0349	-1.63E+02	-0.254556	238
0100	-2.96E+02	-0.489028	430	0350	-1.62E+02	-0.449429	392
0101	-2.95E+02	-0.488683	429	0351	-1.62E+02	-0.439908	383
0102	-2.93E+02	-0.287611	260	0352	-1.61E+02	-0.316927	287
0103	-2.92E+02	-0.636465	482	0353	-1.61E+02	-0.362522	321
0104	-2.92E+02	-0.429213	372	0354	-1.61E+02	-0.084851	111
0105	-2.91E+02	-0.320577	291	0355	-1.61E+02	-0.274242	252
0106	-2.89E+02	-0.499405	437	0356	-1.60E+02	-0.402873	352
0107	-2.89E+02	-0.474482	418	0357	-1.60E+02	0.069228	31
0108	-2.89E+02	-0.139412	147	0358	-1.60E+02	-0.254997	240
0109	-2.89E+02	-0.43118	374	0359	-1.59E+02	-0.131839	140
0110	-2.85E+02	-0.25772	241	0360	-1.59E+02	-0.044011	89
0111	-2.85E+02	-0.319739	290	0361	-1.59E+02	-0.260627	245
0112	-2.84E+02	-0.090058	117	0362	-1.58E+02	-0.384685	334
0113	-2.82E+02	-0.401647	350	0363	-1.58E+02	0.04437	42
0114	-2.80E+02	-0.038839	86	0364	-1.57E+02	-0.301282	271
0115	-2.80E+02	-0.245507	229	0365	-1.57E+02	-0.132037	141
0116	-2.79E+02	-0.295519	266	0366	-1.57E+02	-0.208148	196
0117	-2.78E+02	-0.245406	228	0367	-1.56E+02	-0.398317	349
0118	-2.78E+02	-0.282458	258	0368	-1.56E+02	-0.312556	279
0119	-2.77E+02	-0.295547	267	0369	-1.55E+02	-0.357055	314
0120	-2.75E+02	-0.452845	395	0370	-1.55E+02	-0.552807	458
0121	-2.75E+02	-0.520305	446	0371	-1.55E+02	-0.139333	146
0122	-2.74E+02	0.036367	47	0372	-1.55E+02	-0.266255	247
0123	-2.74E+02	-0.20995	199	0373	-1.55E+02	-0.106077	128
0124	-2.73E+02	-0.442271	389	0374	-1.54E+02	-0.24257	226
0125	-2.72E+02	-0.246318	232	0375	-1.54E+02	0.045177	41
0126	-2.72E+02	-0.385735	337	0376	-1.53E+02	0.071818	28
0127	-2.72E+02	-0.357425	315	0377	-1.53E+02	-0.364421	324
0128	-2.70E+02	-0.431728	375	0378	-1.53E+02	-0.548724	457
0129	-2.70E+02	-0.25047	235	0379	-1.53E+02	-0.218406	205
0130	-2.68E+02	-0.533792	452	0380	-1.52E+02	-0.452905	396
0131	-2.68E+02	-0.627261	479	0381	-1.52E+02	-0.603105	473
0132	-2.68E+02	-0.044107	90	0382	-1.52E+02	-0.163748	159
0133	-2.67E+02	-0.315317	285	0383	-1.51E+02	-0.538615	455

0134	-2.67E+02	-0.124122	136	0384	-1.50E+02	0.195209	6
0135	-2.67E+02	-0.315798	286	0385	-1.49E+02	-0.093961	119
0136	-2.66E+02	-0.155792	154	0386	-1.49E+02	-0.461946	405
0137	-2.66E+02	-0.216479	203	0387	-1.48E+02	-0.102685	123
0138	-2.66E+02	-0.523212	448	0388	-1.48E+02	0.010674	60
0139	-2.65E+02	-0.305066	274	0389	-1.48E+02	-0.441369	385
0140	-2.65E+02	-0.401976	351	0390	-1.48E+02	-0.233028	219
0141	-2.64E+02	-0.413229	360	0391	-1.48E+02	0.013442	58
0142	-2.63E+02	-0.439188	382	0392	-1.47E+02	-0.173404	165
0143	-2.63E+02	-0.463284	409	0393	-1.46E+02	-0.629705	480
0144	-2.63E+02	-0.186051	177	0394	-1.44E+02	-0.433766	378
0145	-2.62E+02	0.103643	23	0395	-1.44E+02	0.018746	57
0146	-2.61E+02	-0.182097	171	0396	-1.44E+02	-0.102769	124
0147	-2.60E+02	-0.374489	329	0397	-1.43E+02	-0.491102	431
0148	-2.60E+02	-0.165479	161	0398	-1.43E+02	0.092314	26
0149	-2.59E+02	0.067772	34	0399	-1.43E+02	-0.277876	254
0150	-2.59E+02	-0.027117	81	0400	-1.42E+02	-0.470821	415
0151	-2.59E+02	-0.235203	220	0401	-1.41E+02	-0.219784	206
0152	-2.57E+02	-0.721584	498	0402	-1.41E+02	-0.173541	166
0153	-2.57E+02	-0.587187	468	0403	-1.41E+02	0.224343	3
0154	-2.56E+02	-0.609968	476	0404	-1.40E+02	-0.05274	94
0155	-2.56E+02	0.037935	45	0405	-1.39E+02	-0.178564	170
0156	-2.55E+02	-0.075161	104	0406	-1.39E+02	-0.717011	497
0157	-2.54E+02	-0.315294	284	0407	-1.39E+02	0.00569	64
0158	-2.53E+02	-0.068132	99	0408	-1.38E+02	0.05572	40
0159	-2.53E+02	-0.442231	388	0409	-1.38E+02	-0.378482	332
0160	-2.53E+02	-0.357774	316	0410	-1.37E+02	-0.146653	150
0161	-2.52E+02	-0.11012	130	0411	-1.37E+02	-0.675045	494
0162	-2.52E+02	-0.531679	451	0412	-1.37E+02	-0.201445	189
0163	-2.52E+02	-0.308752	275	0413	-1.37E+02	-0.231542	218
0164	-2.51E+02	-0.416951	361	0414	-1.36E+02	-0.341312	307
0165	-2.51E+02	0.173492	8	0415	-1.35E+02	-0.603448	474
0166	-2.51E+02	-0.11176	131	0416	-1.35E+02	-0.103071	125
0167	-2.51E+02	0.114002	18	0417	-1.34E+02	0.133262	13
0168	-2.50E+02	0.104418	22	0418	-1.34E+02	-0.254895	239
0169	-2.50E+02	-0.394729	346	0419	-1.34E+02	-0.35705	313
0170	-2.49E+02	-0.377634	331	0420	-1.33E+02	-0.071663	101
0171	-2.49E+02	-0.361756	319	0421	-1.33E+02	0.09967	24
0172	-2.48E+02	-0.224731	213	0422	-1.32E+02	0.282836	2
0173	-2.48E+02	-0.613024	478	0423	-1.32E+02	-0.557361	461
0174	-2.48E+02	-0.268396	250	0424	-1.31E+02	-0.101404	122
0175	-2.48E+02	-0.312781	280	0425	-1.31E+02	-0.278542	256
0176	-2.46E+02	0.064971	35	0426	-1.30E+02	-0.404336	353
0177	-2.46E+02	-0.425222	369	0427	-1.30E+02	-0.640123	484
0178	-2.46E+02	-0.654338	489	0428	-1.29E+02	-0.301006	269
0179	-2.45E+02	-0.03352	84	0429	-1.29E+02	0.064213	36
0180	-2.45E+02	-0.127167	137	0430	-1.29E+02	0.07101	29
0181	-2.45E+02	-0.324463	296	0431	-1.29E+02	-0.228119	216

0182	-2.45E+02	-0.132185	142	0432	-1.29E+02	-0.024933	79
0183	-2.45E+02	-0.407086	355	0433	-1.28E+02	-0.042951	88
0184	-2.45E+02	-0.001421	69	0434	-1.27E+02	0.031667	50
0185	-2.45E+02	-0.385432	336	0435	-1.27E+02	-0.413147	359
0186	-2.44E+02	-0.580212	467	0436	-1.27E+02	-0.087869	113
0187	-2.43E+02	-0.210181	200	0437	-1.25E+02	-0.314578	282
0188	-2.42E+02	-0.386043	338	0438	-1.20E+02	-0.322078	294
0189	-2.42E+02	-0.325685	297	0439	-1.19E+02	0.107388	21
0190	-2.42E+02	-0.052965	95	0440	-1.19E+02	-0.381729	333
0191	-2.42E+02	-0.193626	185	0441	-1.19E+02	-0.511386	443
0192	-2.41E+02	-0.321533	293	0442	-1.18E+02	-0.220796	207
0193	-2.41E+02	-0.032111	83	0443	-1.18E+02	-0.303507	273
0194	-2.40E+02	-0.487138	428	0444	-1.17E+02	-0.134074	144
0195	-2.39E+02	-0.480124	422	0445	-1.17E+02	-0.353567	311
0196	-2.39E+02	-0.519724	445	0446	-1.17E+02	-0.186591	178
0197	-2.39E+02	0.005825	63	0447	-1.17E+02	-0.075363	105
0198	-2.39E+02	-0.178191	169	0448	-1.17E+02	0.091498	27
0199	-2.38E+02	0.000775	67	0449	-1.16E+02	-0.020513	75
0200	-2.37E+02	-0.337451	305	0450	-1.15E+02	-0.308894	276
0201	-2.35E+02	-0.105981	127	0451	-1.14E+02	-0.314543	281
0202	-2.35E+02	0.068161	33	0452	-1.14E+02	-0.421646	366
0203	-2.34E+02	-0.13949	148	0453	-1.14E+02	-0.151097	152
0204	-2.34E+02	-0.455487	401	0454	-1.13E+02	-0.519391	444
0205	-2.34E+02	-0.330942	300	0455	-1.11E+02	-0.096883	121
0206	-2.34E+02	-0.486342	426	0456	-1.11E+02	-0.668208	492
0207	-2.33E+02	0.041146	43	0457	-1.11E+02	-0.260563	244
0208	-2.33E+02	-0.01217	71	0458	-1.10E+02	-0.537949	454
0209	-2.33E+02	-0.117296	134	0459	-1.10E+02	-0.479694	421
0210	-2.33E+02	0.146954	10	0460	-1.10E+02	-0.221787	209
0211	-2.33E+02	-0.045975	92	0461	-1.09E+02	-0.223142	211
0212	-2.32E+02	-0.362622	322	0462	-1.08E+02	-0.081098	109
0213	-2.32E+02	-0.654635	490	0463	-1.08E+02	-0.571441	465
0214	-2.32E+02	-0.483391	423	0464	-1.08E+02	-0.506228	438
0215	-2.31E+02	-0.39399	345	0465	-1.08E+02	-0.137849	145
0216	-2.31E+02	0.062667	37	0466	-1.07E+02	-0.157392	155
0217	-2.31E+02	-0.36529	325	0467	-1.07E+02	0.03174	49
0218	-2.31E+02	-0.441867	386	0468	-1.06E+02	-0.253753	237
0219	-2.30E+02	-0.153571	153	0469	-1.05E+02	-0.240199	223
0220	-2.29E+02	-0.360962	318	0470	-1.05E+02	0.012356	59
0221	-2.29E+02	-0.293246	265	0471	-1.05E+02	-0.421156	364
0222	-2.29E+02	-0.185998	176	0472	-1.05E+02	-0.570491	463
0223	-2.28E+02	-0.230057	217	0473	-1.05E+02	-0.115862	133
0224	-2.28E+02	-0.462873	408	0474	-1.04E+02	-0.088739	114
0225	-2.27E+02	-0.471668	416	0475	-1.04E+02	0.179261	7
0226	-2.27E+02	0.031363	51	0476	-1.03E+02	-0.535507	453
0227	-2.27E+02	-0.201829	192	0477	-1.03E+02	0.057994	39
0228	-2.26E+02	-0.289988	261	0478	-1.02E+02	-0.022929	76
0229	-2.26E+02	-0.411678	358	0479	-1.02E+02	-0.030379	82

0230	-2.25E+02	-0.158098	156	0480	-1.02E+02	-0.246183	231
0231	-2.25E+02	-0.089444	115	0481	-1.02E+02	-0.739348	499
0232	-2.25E+02	-0.389368	343	0482	-1.02E+02	-0.302302	272
0233	-2.24E+02	-0.370262	328	0483	-1.01E+02	-0.332277	301
0234	-2.23E+02	-0.323088	295	0484	-9.97E+01	0.300361	1
0235	-2.23E+02	-0.077329	107	0485	-9.90E+01	-0.004923	70
0236	-2.21E+02	-0.485604	425	0486	-9.88E+01	-0.087279	112
0237	-2.21E+02	0.157318	9	0487	-9.79E+01	-0.4546	399
0238	-2.20E+02	-0.433408	377	0488	-9.77E+01	-0.455837	402
0239	-2.19E+02	-0.216909	204	0489	-9.71E+01	-0.493623	432
0240	-2.19E+02	-0.668843	493	0490	-9.70E+01	-0.207985	195
0241	-2.19E+02	-0.16199	158	0491	-9.65E+01	-0.647774	486
0242	-2.18E+02	-0.186594	179	0492	-9.54E+01	-0.046438	93
0243	-2.18E+02	0.000867	66	0493	-9.49E+01	-0.168668	162
0244	-2.17E+02	-0.571616	466	0494	-9.47E+01	-0.365395	326
0245	-2.17E+02	0.024208	55	0495	-9.40E+01	-0.643009	485
0246	-2.16E+02	-0.253461	236	0496	-9.25E+01	-0.274864	253
0247	-2.13E+02	-0.463662	410	0497	-9.24E+01	0.069443	30
0248	-2.13E+02	0.123689	16	0498	-9.19E+01	-0.12773	138
0249	-2.13E+02	-0.319473	289	0499	-9.19E+01	-0.355973	312
0250	-2.13E+02	-0.037892	85	0500	-9.17E+01	0.006117	62

Table 10-2: Docking Experiment 2/3: Hex docking orientations and STP Scores - The C5/C7 complex

Dock Rank	Energy	T-Score	STP Rank	Dock Rank	Energy	T-Score	STP Rank
0001	-504.6103	-0.469685	417	0251	-142.7896	-0.59952	476
0002	-500.9931	-0.523532	448	0252	-141.1081	-0.28494	264
0003	-454.1128	-0.189172	168	0253	-140.9368	-0.307199	283
0004	-444.6657	-0.004323	54	0254	-140.8616	-0.646522	492
0005	-431.9493	-0.325554	300	0255	-140.4981	-0.285104	265
0006	-415.9352	-0.085634	101	0256	-139.2669	0.067706	23
0007	-409.0476	-0.176959	160	0257	-139.1543	-0.462361	413
0008	-408.54	-0.405089	364	0258	-138.7159	-0.378695	339
0009	-407.4852	-0.291925	269	0259	-137.4623	-0.372647	331
0010	-404.6172	-0.472794	422	0260	-135.452	-0.295889	272
0011	-400.1805	-0.62903	487	0261	-135.2999	-0.577411	472
0012	-397.6649	-0.451436	403	0262	-134.664	-0.111909	118
0013	-395.788	-0.408243	366	0263	-133.9978	-0.127268	128
0014	-388.1381	-0.273874	255	0264	-133.7106	-0.449153	402
0015	-388.1285	-0.012302	59	0265	-132.5105	0.124933	8
0016	-378.1663	-0.475462	424	0266	-132.3565	-0.251994	226
0017	-376.6547	-0.168306	150	0267	-132.0749	-0.303194	279
0018	-375.4805	-0.459565	411	0268	-131.3779	-0.109025	117
0019	-372.493	0.056247	31	0269	-131.3618	0.113018	11
0020	-371.379	-0.57768	473	0270	-127.5251	-0.232235	212

0021	-369.4413	-0.536448	455	0271	-127.1685	0.002013	52
0022	-364.3433	-0.18583	166	0272	-127.1511	-0.491163	431
0023	-361.5594	0.202659	1	0273	-126.962	-0.396084	357
0024	-360.7863	-0.265691	249	0274	-126.3402	-0.527526	451
0025	-358.4532	-0.202336	182	0275	-126.2214	-0.272453	254
0026	-357.7615	0.084761	16	0276	-125.8436	-0.381361	345
0027	-356.4237	-0.386274	349	0277	-125.6946	-0.401795	359
0028	-354.5767	-0.281475	262	0278	-124.6414	-0.600604	478
0029	-353.6708	-0.42813	380	0279	-124.1227	-0.01502	63
0030	-353.4738	-0.021288	68	0280	-123.8399	-0.434299	386
0031	-352.0679	-0.4024	362	0281	-122.4711	-0.139173	134
0032	-349.704	-0.038141	79	0282	-121.6266	-0.046458	81
0033	-349.3951	-0.534675	454	0283	-121.4287	-0.266513	250
0034	-348.4401	0.039198	38	0284	-121.2553	-0.145162	136
0035	-346.0467	-0.452934	404	0285	-120.7341	0.042856	33
0036	-345.6039	-0.453175	405	0286	-119.5112	0.035291	40
0037	-344.4354	-0.508251	440	0287	-118.3967	-0.128717	131
0038	-342.6106	-0.306997	282	0288	-117.8643	0.060151	28
0039	-342.1909	-0.124379	125	0289	-116.4594	-0.336937	307
0040	-340.7101	-0.618983	484	0290	-113.1032	0.09445	15
0041	-337.2875	-0.706253	499	0291	-112.8185	-0.082348	99
0042	-336.9212	-0.009147	56	0292	-112.4134	0.079893	20
0043	-334.6778	-0.262373	240	0293	-112.1282	-0.013971	61
0044	-326.3958	-0.161178	144	0294	-109.8185	-0.024985	72
0045	-326.1649	-0.539289	456	0295	-109.5952	-0.280084	260
0046	-324.3026	-0.207742	188	0296	-109.4517	-0.342402	315
0047	-324.261	-0.563035	464	0297	-109.2049	-0.216758	198
0048	-324.0058	-0.438228	394	0298	-108.9616	-0.208418	189
0049	-321.1642	-0.237008	218	0299	-108.0184	-0.137849	133
0050	-318.7661	-0.355871	324	0300	-107.7106	-0.279477	259
0051	-317.1318	-0.417829	372	0301	-106.1584	-0.181532	163
0052	-316.2595	-0.480667	427	0302	-106.1475	-0.426596	377
0053	-316.2291	-0.194285	172	0303	-105.1862	-0.266628	251
0054	-315.2596	-0.297482	275	0304	-104.7386	-0.118506	121
0055	-314.7564	-0.469952	418	0305	-104.3429	-0.198317	177
0056	-313.9455	-0.457013	409	0306	-104.0134	-0.407239	365
0057	-309.2409	-0.473886	423	0307	-103.6835	-0.210928	195
0058	-306.1189	-0.125934	126	0308	-102.3145	-0.023402	69
0059	-305.3232	-0.012037	58	0309	-101.8492	-0.524776	449
0060	-302.6103	-0.201568	181	0310	-101.3454	-0.378734	340
0061	-302.5136	-0.338523	309	0311	-99.19016	-0.571285	470
0062	-301.8129	-0.328895	302	0312	-98.65167	-0.068421	91
0063	-300.512	-0.206514	186	0313	-97.37617	-0.355027	323
0064	-300.176	-0.375258	336	0314	-93.68138	-0.527063	450
0065	-299.5643	-0.456795	408	0315	-93.34427	-0.339793	310
0066	-297.2619	-0.160934	143	0316	-93.25174	-0.516355	444
0067	-294.2411	-0.146138	137	0317	-92.09692	-0.618422	483
0068	-293.9784	-0.493036	434	0318	-92.09634	-0.015026	64

0069	-293.8652	-0.066758	90	0319	-91.83072	-0.34961	320
0070	-293.4175	-0.454553	406	0320	-91.5455	-0.118097	119
0071	-292.7534	-0.102064	113	0321	-91.43686	-0.588953	474
0072	-292.1592	-0.402484	363	0322	-90.09546	0.030488	42
0073	-289.6729	-0.313825	285	0323	-89.51842	-0.413165	370
0074	-288.8351	0.144842	4	0324	-89.49051	-0.127276	129
0075	-288.5347	-0.465982	415	0325	-89.2132	-0.096628	111
0076	-287.6547	-0.509059	441	0326	-88.93826	-0.21178	196
0077	-286.7291	0.040655	36	0327	-88.33566	-0.256578	233
0078	-283.983	-0.420359	374	0328	-87.29449	-0.430653	382
0079	-283.3239	-0.208936	190	0329	-85.75256	-0.174877	157
0080	-283.1609	-0.446052	401	0330	-85.1351	-0.445836	400
0081	-278.0793	-0.26344	245	0331	-84.28545	-0.203896	183
0082	-276.2859	0.106063	12	0332	-84.05121	-0.400806	358
0083	-274.5035	0.126742	7	0333	-82.6687	-0.569846	469
0084	-274.2173	-0.317754	291	0334	-82.25089	-0.295267	271
0085	-269.9803	-0.471225	420	0335	-80.69162	-0.066225	89
0086	-266.1445	-0.33209	305	0336	-80.06104	-0.252032	227
0087	-265.5003	-0.144494	135	0337	-79.59476	-0.605833	480
0088	-264.9919	-0.180589	162	0338	-76.52039	-0.169086	152
0089	-262.8959	-0.091323	106	0339	-75.54129	0.02913	44
0090	-260.2406	-0.470655	419	0340	-75.40067	-0.331013	303
0091	-259.9178	-0.554587	463	0341	-75.3938	-0.231622	211
0092	-258.993	-0.17331	155	0342	-75.17906	-0.435733	388
0093	-258.9346	-0.258045	235	0343	-75.04907	-0.431736	383
0094	-258.4725	-0.441	395	0344	-75.03513	-0.493155	435
0095	-258.4103	-0.195643	174	0345	-74.7258	-0.210327	193
0096	-257.6059	-0.380224	341	0346	-74.5589	-0.282833	263
0097	-257.1373	-0.38941	353	0347	-73.44757	-0.324444	296
0098	-256.2792	-0.002197	53	0348	-71.98959	-0.020153	66
0099	-256.0731	-0.163788	147	0349	-71.3521	-0.338074	308
0100	-253.6316	-0.563046	465	0350	-70.50706	-0.376036	337
0101	-253.1773	-0.442231	397	0351	-69.19608	-0.317011	289
0102	-253.0374	-0.280996	261	0352	-68.69286	0.067408	24
0103	-252.7698	-0.262168	239	0353	-68.05072	-0.516623	445
0104	-252.1243	-0.44246	398	0354	-67.92822	-0.11901	122
0105	-250.7114	-0.209962	191	0355	-67.92599	-0.028503	75
0106	-250.1457	-0.364747	325	0356	-66.91653	-0.027628	74
0107	-249.1453	-0.148776	139	0357	-65.10886	0.06423	25
0108	-248.6688	-0.428129	379	0358	-63.90811	-0.29693	273
0109	-248.6366	-0.613024	481	0359	-63.60818	-0.228678	208
0110	-246.2385	0.138917	5	0360	-63.446	-0.036861	78
0111	-243.696	0.133653	6	0361	-63.40102	-0.200162	178
0112	-242.4472	0.08256	19	0362	-63.31554	-0.231366	210
0113	-242.3637	-0.210181	192	0363	-63.30676	0.057231	30
0114	-241.1946	-0.663949	493	0364	-63.01007	0.041907	35
0115	-240.535	-0.395163	356	0365	-62.72824	-0.457405	410
0116	-240.0072	0.162344	3	0366	-62.40643	-0.354834	322

0117	-239.8235	-0.29112	266	0367	-61.9567	-0.719456	500
0118	-237.3213	-0.533824	452	0368	-60.80164	-0.424593	376
0119	-236.2196	-0.096145	110	0369	-60.60999	-0.118324	120
0120	-236.2027	0.104818	13	0370	-59.60878	-0.258809	237
0121	-233.331	-0.069303	92	0371	-56.13342	0.034347	41
0122	-233.0334	-0.51573	443	0372	-56.01328	-0.267754	252
0123	-230.3772	-0.010912	57	0373	-55.99623	-0.388313	352
0124	-229.4106	-0.235966	217	0374	-54.93863	-0.049215	84
0125	-228.5817	-0.56724	466	0375	-54.1788	-0.642265	491
0126	-228.5656	-0.365415	326	0376	-54.14505	-0.623903	485
0127	-227.4581	-0.251451	225	0377	-54.0943	-0.401833	360
0128	-224.0882	-0.432232	385	0378	-53.26703	-0.047201	82
0129	-222.3122	-0.255033	230	0379	-52.72626	-0.697579	498
0130	-222.1864	-0.314208	286	0380	-52.51309	-0.29192	268
0131	-220.9965	-0.262699	243	0381	-52.33066	-0.391467	354
0132	-220.1712	-0.190215	170	0382	-52.25516	-0.23231	213
0133	-218.3684	-0.056536	85	0383	-50.11902	-0.194406	173
0134	-218.2728	-0.376057	338	0384	-49.31873	-0.223381	202
0135	-218.1973	-0.554537	462	0385	-48.9769	0.164615	2
0136	-217.9038	-0.245753	222	0386	-48.62637	-0.443658	399
0137	-217.8102	-0.343557	317	0387	-48.56204	-0.189539	169
0138	-217.3677	-0.162063	145	0388	-48.22763	-0.206892	187
0139	-217.3528	-0.235293	215	0389	-47.80963	-0.368866	327
0140	-217.0495	-0.176316	159	0390	-46.00711	-0.153855	141
0141	-216.9437	-0.387883	350	0391	-45.8317	-0.262536	241
0142	-216.5728	-0.297186	274	0392	-45.3049	-0.212986	197
0143	-216.5457	-0.264597	247	0393	-44.51523	-0.625656	486
0144	-216.3423	-0.006514	55	0394	-43.48166	-0.59531	475
0145	-215.4614	-0.629475	488	0395	-43.21613	-0.169749	153
0146	-215.2834	-0.298217	276	0396	-42.95151	-0.345843	319
0147	-215.2675	-0.551308	461	0397	-42.49725	-0.173687	156
0148	-214.8392	-0.210663	194	0398	-42.31314	-0.381813	346
0149	-214.8091	-0.122153	123	0399	-40.87503	-0.423342	375
0150	-214.4418	-0.548094	459	0400	-39.1051	-0.04786	83
0151	-213.9971	-0.341348	312	0401	-38.00055	-0.220842	200
0152	-212.5535	0.013335	47	0402	-37.71652	-0.304974	280
0153	-212.3757	-0.076819	96	0403	-37.10843	-0.220947	201
0154	-211.9959	-0.089264	104	0404	-35.12952	-0.325006	297
0155	-211.8115	-0.686278	496	0405	-35.10858	-0.437249	392
0156	-211.431	-0.061668	87	0406	-34.96252	-0.257537	234
0157	-210.0984	0.011706	48	0407	-31.69708	-0.132907	132
0158	-209.5864	-0.253264	229	0408	-29.66602	-0.471765	421
0159	-209.0928	-0.197855	176	0409	-28.05289	-0.325415	299
0160	-209.0415	-0.480003	425	0410	-27.87505	-0.322629	295
0161	-207.9008	-0.027072	73	0411	-27.48154	-0.226412	204
0162	-207.6702	-0.520991	447	0412	-27.20013	-0.264303	246
0163	-207.3615	-0.200388	179	0413	-26.24738	-0.146705	138
0164	-206.632	-0.255364	231	0414	-26.23636	-0.02107	67

0165	-205.9945	0.03555	39	0415	-25.00558	-0.550721	460
0166	-205.14	-0.320337	292	0416	-24.66904	-0.071082	93
0167	-204.2464	-0.408318	367	0417	-23.66727	0.123079	9
0168	-202.8697	-0.417121	371	0418	-23.32614	0.002383	51
0169	-202.6793	-0.04314	80	0419	-22.62216	-0.014856	62
0170	-202.3872	-0.682238	494	0420	-21.43626	-0.299809	277
0171	-200.2219	-0.16572	148	0421	-21.05209	-0.316669	287
0172	-198.4733	-0.480172	426	0422	-20.4285	-0.260026	238
0173	-197.2204	-0.228081	206	0423	-19.52936	-0.484438	429
0174	-196.9576	-0.374457	333	0424	-17.90567	-0.380349	342
0175	-196.8728	-0.340752	311	0425	-17.74695	0.053218	32
0176	-196.007	0.068721	22	0426	-17.70853	0.057546	29
0177	-193.0244	-0.317006	288	0427	-17.07407	-0.087432	103
0178	-191.8773	-0.094175	109	0428	-16.67404	-0.507105	439
0179	-191.772	-0.275263	256	0429	-16.19235	-0.373933	332
0180	-191.4909	-0.206213	185	0430	-15.06723	-0.391975	355
0181	-191.1862	-0.184035	165	0431	-13.91629	-0.218842	199
0182	-190.6449	-0.27088	253	0432	-12.42877	-0.128675	130
0183	-188.6929	-0.263137	244	0433	-11.67413	-0.124268	124
0184	-187.1242	-0.320682	294	0434	-10.42944	-0.383488	347
0185	-185.2944	0.062211	26	0435	-6.819489	-0.091637	107
0186	-183.8522	-0.436388	389	0436	-6.656708	0.069316	21
0187	-183.4828	-0.385671	348	0437	-6.562286	-0.030845	76
0188	-182.9055	-0.491855	432	0438	-6.472626	-0.291753	267
0189	-182.605	-0.511566	442	0439	-6.090332	0.040556	37
0190	-182.1108	-0.325308	298	0440	-5.396515	0.100323	14
0191	-182.1074	-0.204346	184	0441	-5.250702	-0.306042	281
0192	-182.1053	-0.498805	437	0442	-4.90741	-0.092563	108
0193	-181.9908	-0.491877	433	0443	-4.703339	-0.352514	321
0194	-181.7951	-0.693457	497	0444	-4.16922	-0.33268	306
0195	-179.2631	-0.343719	318	0445	-2.565094	-0.086209	102
0196	-179.1186	-0.179475	161	0446	-2.071259	-0.105126	115
0197	-178.9259	-0.056827	86	0447	-1.879745	-0.151732	140
0198	-178.4267	-0.429531	381	0448	-1.079529	-0.10559	116
0199	-178.3694	0.117041	10	0449	-0.7724609	-0.276133	258
0200	-177.4827	-0.46374	414	0450	-0.6037292	-0.081881	98
0201	-175.1204	-0.25573	232	0451	-0.4438171	-0.165725	149
0202	-173.8738	-0.412833	369	0452	0.4293823	-0.59965	477
0203	-173.7692	-0.240994	219	0453	1.510956	-0.418759	373
0204	-173.557	-0.012904	60	0454	2.795105	0.060644	27
0205	-172.9586	-0.226079	203	0455	3.496262	-0.12721	127
0206	-172.4937	-0.547668	458	0456	4.581665	-0.192529	171
0207	-172.3316	-0.519985	446	0457	4.782013	-0.30073	278
0208	-172.0797	0.041955	34	0458	5.604706	-0.024106	70
0209	-170.9683	-0.495628	436	0459	5.953369	-0.684106	495
0210	-170.0472	-0.534013	453	0460	6.294647	-0.442115	396
0211	-169.758	-0.374825	334	0461	6.483368	0.083582	18
0212	-169.2398	-0.228155	207	0462	6.945526	-0.262645	242

0213	-169.1465	-0.500355	438	0463	7.534973	-0.275608	257
0214	-168.5713	-0.455583	407	0464	7.804413	-0.48144	428
0215	-167.8845	-0.432169	384	0465	8.487427	-0.241993	220
0216	-167.4384	-0.614536	482	0466	8.584808	-0.182283	164
0217	-165.9222	-0.341627	313	0467	9.205811	-0.436698	391
0218	-165.63	-0.317199	290	0468	9.444733	0.010434	49
0219	-164.6217	-0.018123	65	0469	9.615448	-0.40217	361
0220	-163.8657	-0.46733	416	0470	9.923004	-0.234942	214
0221	-163.0727	-0.331843	304	0471	10.95761	0.020965	45
0222	-162.7193	-0.56821	467	0472	11.62714	-0.080407	97
0223	-162.3245	-0.074047	95	0473	12.61218	-0.375244	335
0224	-160.9387	-0.097673	112	0474	13.1676	-0.605118	479
0225	-160.1148	-0.343487	316	0475	13.31119	-0.544305	457
0226	-159.7282	-0.388223	351	0476	13.75323	-0.435668	387
0227	-159.6455	-0.371167	329	0477	13.79929	-0.568399	468
0228	-158.8546	-0.226773	205	0478	14.95544	-0.29512	270
0229	-158.5439	-0.235537	216	0479	15.35516	-0.082856	100
0230	-157.5409	0.030103	43	0480	15.4075	-0.036018	77
0231	-155.6985	-0.09119	105	0481	15.5499	-0.327662	301
0232	-155.3886	-0.264638	248	0482	15.64441	-0.175943	158
0233	-154.2484	0.083921	17	0483	15.67122	-0.320362	293
0234	-153.6442	-0.427219	378	0484	16.02695	-0.248465	224
0235	-152.4896	-0.312113	284	0485	16.10376	-0.200475	180
0236	-152.2875	-0.380626	343	0486	16.26764	-0.370165	328
0237	-149.8917	-0.371543	330	0487	17.77176	-0.15942	142
0238	-147.6485	-0.07364	94	0488	18.08006	-0.242154	221
0239	-147.6401	-0.024295	71	0489	18.25888	-0.187701	167
0240	-147.34	-0.380744	344	0490	18.59744	-0.41206	368
0241	-146.2582	-0.460203	412	0491	19.78607	-0.486913	430
0242	-145.8578	-0.436604	390	0492	21.16513	-0.168752	151
0243	-145.7707	-0.631385	489	0493	21.98569	-0.102355	114
0244	-145.3773	-0.062934	88	0494	22.06555	-0.252594	228
0245	-144.6088	-0.162649	146	0495	22.72794	0.020391	46
0246	-144.4216	-0.247786	223	0496	23.25388	-0.641605	490
0247	-144.1501	-0.173045	154	0497	23.75058	-0.197637	175
0248	-144.0187	-0.342064	314	0498	24.44489	-0.437637	393
0249	-143.4838	-0.230063	209	0499	24.74133	0.008937	50
0250	-143.0339	-0.258397	236	0500	25.9325	-0.577105	471

Table 10-3: Docking Experiment 3/3: Hex docking orientations and STP Scores - The C5/C7 complex

Dock Rank	Energy	T-Score	STP Rank	Dock Rank	Energy	T-Score	STP Rank
0001	-766.0867	-0.469685	418	0251	-549.9763	-0.37235	339
0002	-740.6084	-0.226079	218	0252	-549.6996	-0.476929	425
0003	-731.8795	-0.211334	205	0253	-549.3289	-0.447317	405
0004	-726.4556	-0.541441	457	0254	-549.2297	-0.032526	64
0005	-702.5189	-0.233606	231	0255	-549.1938	-0.314208	300
0006	-697.2649	-0.011223	45	0256	-549.1311	-0.098586	110
0007	-694.2816	0.004534	37	0257	-549.0038	-0.021339	54
0008	-691.8359	-0.342564	313	0258	-548.8706	-0.455135	410
0009	-691.1224	-0.108699	118	0259	-547.9839	-0.377105	345
0010	-685.4019	-0.245886	240	0260	-547.9072	-0.07364	93
0011	-681.8054	-0.076819	99	0261	-547.7188	-0.382548	357
0012	-676.9677	-0.024985	58	0262	-547.1843	-0.350751	319
0013	-674.2859	0.050728	23	0263	-547.1493	-0.361738	328
0014	-671.8557	-0.527575	449	0264	-547.1171	-0.364885	330
0015	-670.9099	-0.219761	212	0265	-547.0234	-0.128702	137
0016	-662.6126	-0.495628	435	0266	-546.8765	-0.165725	165
0017	-660.6943	-0.07636	97	0267	-546.6829	0.072236	14
0018	-660.2383	-0.587014	480	0268	-546.5867	0.057231	20
0019	-657.8298	-0.162649	163	0269	-546.5338	-0.519985	443
0020	-657.6893	-0.139431	149	0270	-546.4947	-0.026837	59
0021	-655.1075	-0.605833	485	0271	-546.4065	-0.37569	342
0022	-651.4045	-0.429531	390	0272	-546.1423	-0.118484	127
0023	-648.4456	-0.231882	227	0273	-546.0488	-0.173045	173
0024	-647.9976	-0.198358	192	0274	-546.0151	-0.351832	321
0025	-644.5079	-0.110519	119	0275	-545.6505	-0.299809	291
0026	-641.3528	-0.261598	259	0276	-545.507	-0.041248	68
0027	-639.4129	-0.327713	308	0277	-545.5049	-0.451436	407
0028	-637.8615	-0.27088	270	0278	-545.4434	-0.541031	456
0029	-633.6868	-0.249414	242	0279	-545.4189	-0.190215	186
0030	-632.4712	0.068908	16	0280	-545.3174	-0.384808	359
0031	-630.1116	-0.385671	360	0281	-544.8198	-0.033174	65
0032	-629.8683	-0.071706	92	0282	-544.4868	-0.533864	452
0033	-628.947	-0.177233	177	0283	-544.4159	-0.199217	193
0034	-627.6917	-0.645446	495	0284	-544.3425	-0.256764	249
0035	-627.4977	-0.650622	496	0285	-544.0597	-0.417829	384
0036	-627.0496	-0.157822	158	0286	-543.7377	-0.113195	124
0037	-626.6765	0.030488	32	0287	-543.7101	-0.456795	412
0038	-625.4041	-0.169471	169	0288	-543.1393	-0.380744	352
0039	-624.4987	-0.472706	423	0289	-543.1013	-0.491855	432
0040	-624.4441	-0.256578	248	0290	-542.9018	-0.117376	125
0041	-623.2947	0.007997	36	0291	-542.8048	-0.550403	463
0042	-622.3146	-0.353213	323	0292	-542.5329	-0.058637	85
0043	-619.8041	-0.206213	199	0293	-542.4651	-0.370024	334
0044	-619.3777	-0.401833	372	0294	-542.3655	-0.431736	393

0045	-619.2737	-0.169086	168	0295	-542.2941	-0.146138	152
0046	-617.9756	-0.315903	301	0296	-542.1401	-0.537386	454
0047	-617.9578	-0.391903	367	0297	-541.7997	-0.096549	107
0048	-617.0548	-0.084293	102	0298	-541.7055	-0.426735	388
0049	-616.8309	-0.346091	314	0299	-541.6934	-0.104636	115
0050	-615.28	-0.049341	73	0300	-541.6262	0.032954	30
0051	-614.7875	-0.663949	498	0301	-541.613	-0.231366	225
0052	-614.6195	-0.119733	130	0302	-541.5563	-0.258397	253
0053	-613.6375	-0.085634	103	0303	-541.4449	-0.538323	455
0054	-612.6393	-0.442115	400	0304	-541.2081	0.010434	34
0055	-612.6279	-0.200475	195	0305	-541.169	-0.282458	274
0056	-612.026	-0.40217	373	0306	-540.7716	0.062211	19
0057	-611.9865	-0.449642	406	0307	-540.7279	-0.084247	101
0058	-611.6705	-0.266628	267	0308	-540.7054	-0.260537	256
0059	-610.266	-0.170509	171	0309	-540.6531	-0.436698	397
0060	-608.4976	-0.617154	486	0310	-540.5775	-0.151642	154
0061	-608.4404	-0.347996	316	0311	-540.4515	-0.388223	362
0062	-607.5881	-0.056805	83	0312	-540.2646	-0.619429	488
0063	-605.8177	-0.59965	482	0313	-540.0677	-0.178431	178
0064	-605.3702	-0.05445	81	0314	-539.9854	0.077274	13
0065	-605.1031	-0.156091	157	0315	-539.9302	-0.33268	310
0066	-603.6841	-0.376036	343	0316	-539.5663	-0.05049	75
0067	-602.3535	-0.378695	346	0317	-539.4515	0.108021	5
0068	-602.153	-0.542536	460	0318	-539.3547	-0.132324	141
0069	-601.8796	-0.189117	183	0319	-538.8293	-0.380224	349
0070	-600.8924	-0.382307	355	0320	-538.5748	-0.233432	230
0071	-599.4274	-0.267565	268	0321	-538.5737	-0.075907	96
0072	-598.3375	-0.10415	113	0322	-538.2804	-0.022214	55
0073	-598.3225	-0.554481	466	0323	-538.2298	-0.319405	303
0074	-597.1606	-0.285	275	0324	-538.1836	-0.382413	356
0075	-595.1707	-0.136212	146	0325	-538.1515	-0.447064	404
0076	-595.1646	-0.19355	189	0326	-537.9052	-0.599698	483
0077	-594.7185	-0.232197	229	0327	-537.8159	-0.470443	421
0078	-594.625	-0.002197	39	0328	-537.7211	-0.502144	438
0079	-592.3729	-0.042855	69	0329	-537.5299	-0.132907	142
0080	-592.2362	-0.11262	123	0330	-537.2479	-0.099148	111
0081	-592.0492	-0.29512	286	0331	-537.0412	-0.44276	401
0082	-591.6574	-0.523208	444	0332	-536.9045	-0.223381	216
0083	-590.9056	-0.133076	143	0333	-536.8974	-0.12558	132
0084	-590.8661	-0.290567	282	0334	-536.8142	-0.045987	70
0085	-590.8029	-0.258548	254	0335	-536.803	-0.313825	299
0086	-590.6509	-0.523532	445	0336	-536.6445	-0.434299	394
0087	-590.6213	-0.11195	122	0337	-536.5566	-0.565143	472
0088	-590.5206	-0.391083	366	0338	-536.5337	-0.028589	62
0089	-590.3292	-0.371167	337	0339	-536.5086	-0.542251	459
0090	-589.8795	-0.352514	322	0340	-536.2573	-0.207764	202
0091	-589.2017	-0.512357	440	0341	-536.1933	-0.307199	295
0092	-588.6682	-0.471765	422	0342	-536.1528	-0.027927	61

0093	-587.8341	-0.194285	190	0343	-536.1523	-0.369502	332
0094	-586.8952	-0.118118	126	0344	-535.8919	-0.243929	239
0095	-586.7219	-0.3802	348	0345	-535.8451	-0.259166	255
0096	-586.3894	-0.4024	374	0346	-535.6922	-0.554587	468
0097	-586.3838	-0.013479	48	0347	-535.4654	-0.021288	53
0098	-586.2693	-0.444441	402	0348	-535.1996	-0.406094	376
0099	-586.1012	-0.634332	492	0349	-535.1893	-0.105126	116
0100	-585.4354	-0.012302	46	0350	-535.1073	-0.271016	271
0101	-585.2228	-0.074652	95	0351	-534.9594	-0.469894	420
0102	-585.2043	-0.264638	264	0352	-534.9098	-0.292076	285
0103	-584.4926	-0.053511	79	0353	-534.7615	0.104818	7
0104	-584.3892	-0.227851	220	0354	-534.619	-0.405089	375
0105	-584.3699	-0.493005	434	0355	-534.4962	-0.408577	379
0106	-584.1812	-0.486699	429	0356	-534.485	-0.126056	135
0107	-582.9926	-0.556737	469	0357	-534.4654	-0.070327	91
0108	-582.9917	-0.39083	365	0358	-534.2651	-0.351156	320
0109	-582.9414	-0.577411	477	0359	-534.1807	-0.118606	128
0110	-582.8251	-0.2639	262	0360	-534.0885	-0.291953	284
0111	-582.548	-0.08138	100	0361	-534.0723	-0.028908	63
0112	-582.4847	-0.497797	436	0362	-533.9254	-0.388898	363
0113	-581.7062	-0.209394	203	0363	-533.9052	-0.640999	494
0114	-580.5038	-0.193289	188	0364	-533.7382	0.038545	27
0115	-580.4958	0.079893	12	0365	-533.6844	-0.697579	500
0116	-580.465	-0.380622	350	0366	-533.527	0.034347	29
0117	-580.2401	-0.453416	408	0367	-533.5242	0.102789	8
0118	-579.6302	-0.168463	167	0368	-533.5226	-0.11946	129
0119	-579.5398	-0.107797	117	0369	-533.4764	-0.076362	98
0120	-579.4927	-0.36161	327	0370	-533.472	-0.320337	304
0121	-579.0681	-0.26147	258	0371	-533.2686	-0.327662	307
0122	-579.0214	-0.655106	497	0372	-533.198	-0.515811	442
0123	-578.3697	-0.371543	338	0373	-533.0596	-0.406604	377
0124	-578.1146	-0.428129	389	0374	-532.988	-0.441	398
0125	-577.9903	-0.416169	383	0375	-532.9864	-0.357647	326
0126	-577.6519	-0.018123	51	0376	-532.9553	-0.469293	417
0127	-577.5919	-0.29192	283	0377	-532.9025	-0.527063	448
0128	-577.5727	-0.187701	182	0378	-532.3351	-0.206514	200
0129	-577.5371	-0.223201	215	0379	-532.2955	0.04517	25
0130	-577.4365	-0.682238	499	0380	-532.2266	-0.263137	261
0131	-576.9731	-0.21178	206	0381	-532.1602	-0.062623	88
0132	-576.8979	-0.370225	336	0382	-532.1298	-0.024295	57
0133	-576.2949	-0.050639	76	0383	-532.075	-0.015538	49
0134	-576.2523	-0.131923	140	0384	-532.026	-0.390011	364
0135	-576.2362	-0.541561	458	0385	-531.9685	-0.201361	197
0136	-575.063	-0.441731	399	0386	-531.8843	-0.473886	424
0137	-575.0546	0.043975	26	0387	-531.8011	-0.554537	467
0138	-574.6344	-0.526927	447	0388	-531.7665	-0.229848	224
0139	-574.6091	-0.465193	416	0389	-531.7604	-0.462929	414
0140	-573.3682	-0.024106	56	0390	-531.6893	-0.046458	71

0141	-572.2736	-0.227946	222	0391	-531.6847	-0.08761	105
0142	-571.9967	0.1574	2	0392	-531.5937	-0.131135	139
0143	-571.6605	-0.563046	471	0393	-531.5163	-0.364806	329
0144	-571.4863	-0.047201	72	0394	-531.4826	-0.275263	272
0145	-570.8168	-0.262746	260	0395	-531.4721	-0.514036	441
0146	-570.7456	-0.035661	66	0396	-531.3361	-0.544679	461
0147	-570.4425	-0.387883	361	0397	-531.3275	0.038337	28
0148	-570.386	0.084761	11	0398	-531.2601	-0.257537	250
0149	-570.3835	-0.24948	243	0399	-531.246	-0.407178	378
0150	-569.8262	-0.430653	391	0400	-531.2162	-0.009147	42
0151	-569.2159	-0.052745	78	0401	-531.1621	-0.21587	210
0152	-568.9056	-0.176959	176	0402	-530.78	-0.559014	470
0153	-568.7099	-0.372849	340	0403	-530.6364	-0.455751	411
0154	-568.6628	-0.141546	150	0404	-530.5178	-0.550721	464
0155	-568.5701	-0.297186	290	0405	-530.4958	-0.111341	120
0156	-567.9236	-0.380626	351	0406	-530.4467	-0.381877	354
0157	-567.3312	-0.326864	306	0407	-530.436	-0.285575	276
0158	-567.042	-0.276133	273	0408	-530.3577	-0.400036	370
0159	-566.9911	-0.166393	166	0409	-530.3039	-0.135316	145
0160	-566.7155	-0.423633	387	0410	-530.2609	-0.265691	265
0161	-566.6993	-0.585769	479	0411	-530.1627	-0.051453	77
0162	-566.3927	-0.124268	131	0412	-529.7803	-0.059738	86
0163	-566.0929	-0.434356	395	0413	-529.7397	-0.286347	278
0164	-565.9316	-0.290164	281	0414	-529.5591	-0.184035	180
0165	-565.914	0.053615	22	0415	-529.3138	-0.266513	266
0166	-565.6263	0.002013	38	0416	-529.308	0.056247	21
0167	-565.3893	-0.017934	50	0417	-529.2769	-0.444586	403
0168	-564.4758	-0.210928	204	0418	-529.1495	-0.376338	344
0169	-564.4003	-0.419646	386	0419	-529.0891	-0.213055	207
0170	-563.7557	-0.248969	241	0420	-528.9257	-0.252331	244
0171	-563.6241	-0.097534	108	0421	-528.8872	-0.302197	292
0172	-563.369	-0.62918	490	0422	-528.8327	-0.161178	161
0173	-563.3377	-0.629475	491	0423	-528.7274	-0.010631	44
0174	-563.3182	-0.252594	245	0424	-528.7106	-0.237512	235
0175	-563.1607	0.067408	18	0425	-528.7087	-0.192529	187
0176	-562.8753	-0.529615	450	0426	-528.6181	-0.050384	74
0177	-562.5423	-0.534013	453	0427	-528.5583	-0.296927	289
0178	-562.2987	-0.296463	288	0428	-528.5568	-0.200388	194
0179	-561.8029	-0.384773	358	0429	-528.4867	-0.469892	419
0180	-561.2847	-0.036018	67	0430	-528.3792	-0.480172	426
0181	-561.2117	-0.220549	213	0431	-528.377	-0.620202	489
0182	-560.9216	-0.436604	396	0432	-528.3691	-0.483574	427
0183	-560.2297	-0.232095	228	0433	-528.2943	-0.56724	473
0184	-560.135	-0.138203	148	0434	-528.195	-0.003825	40
0185	-560.09	0.069553	15	0435	-528.0976	-0.154738	155
0186	-559.7935	-0.346714	315	0436	-528.0747	-0.05444	80
0187	-559.7074	-0.255597	247	0437	-527.6584	-0.22463	217
0188	-559.5187	-0.141657	151	0438	-527.5247	-0.334521	311

0189	-559.2996	-0.16572	164	0439	-527.4914	-0.551593	465
0190	-559.0032	-0.484438	428	0440	-527.438	-0.309231	296
0191	-558.6926	-0.17014	170	0441	-527.3522	-0.459984	413
0192	-558.6877	-0.391975	368	0442	-527.2683	-0.138167	147
0193	-558.6691	-0.179475	179	0443	-527.1702	-0.253729	246
0194	-558.5934	-0.571467	476	0444	-527.0902	-0.4133	381
0195	-558.3807	-0.353499	324	0445	-527.0411	-0.087152	104
0196	-558.369	-0.463005	415	0446	-527.0154	-0.295178	287
0197	-558.2168	-0.378734	347	0447	-526.9824	-0.410578	380
0198	-558.0954	0.008937	35	0448	-526.9576	-0.331276	309
0199	-557.8243	-0.213466	209	0449	-526.9434	0.202659	1
0200	-557.7941	-0.227102	219	0450	-526.9139	-0.111909	121
0201	-557.7027	-0.189539	185	0451	-526.802	-0.401795	371
0202	-557.5487	-0.3701	335	0452	-526.7551	-0.640776	493
0203	-557.3466	-0.48897	431	0453	-526.736	0.049062	24
0204	-557.2495	-0.14835	153	0454	-526.6122	-0.026955	60
0205	-556.8574	-0.20301	198	0455	-526.5735	-0.068421	90
0206	-556.7777	-0.213261	208	0456	-526.5605	-0.353616	325
0207	-556.4247	-0.004323	41	0457	-526.5593	-0.603957	484
0208	-556.2839	-0.130643	138	0458	-526.3725	-0.125607	133
0209	-556.193	-0.287885	279	0459	-526.3151	-0.366182	331
0210	-556.0766	-0.594503	481	0460	-526.0862	-0.201131	196
0211	-555.9524	-0.010217	43	0461	-525.8876	0.09445	9
0212	-555.8376	-0.02107	52	0462	-525.7997	-0.532156	451
0213	-555.8362	-0.01332	47	0463	-525.7531	-0.160397	160
0214	-555.7247	-0.102355	112	0464	-525.677	-0.414714	382
0215	-555.6307	-0.062383	87	0465	-525.4865	-0.160347	159
0216	-555.413	-0.316669	302	0466	-525.4796	-0.264636	263
0217	-555.2244	-0.154789	156	0467	-525.4575	0.123079	4
0218	-555.0236	-0.187407	181	0468	-525.3124	-0.375244	341
0219	-554.8707	-0.418759	385	0469	-525.264	-0.320682	305
0220	-554.8348	-0.492456	433	0470	-525.1871	-0.306086	294
0221	-554.2554	-0.311383	297	0471	-525.1181	-0.062934	89
0222	-553.6578	-0.392447	369	0472	-524.9631	-0.454531	409
0223	-553.5702	-0.431423	392	0473	-524.9269	-0.57768	478
0224	-553.2659	-0.260823	257	0474	-524.8506	-0.511635	439
0225	-552.8698	-0.348996	318	0475	-524.8403	-0.135252	144
0226	-552.8036	-0.348813	317	0476	-524.7929	-0.381299	353
0227	-552.7965	-0.233756	232	0477	-524.7637	-0.498805	437
0228	-552.6351	0.031389	31	0478	-524.6945	-0.175322	175
0229	-552.6207	0.126897	3	0479	-524.64	-0.242569	238
0230	-552.6035	-0.241993	237	0480	-524.6274	-0.369597	333
0231	-552.5973	-0.526381	446	0481	-524.4711	-0.270808	269
0232	-552.4307	-0.28624	277	0482	-524.4233	-0.288956	280
0233	-552.3334	-0.257538	251	0483	-524.405	-0.127268	136
0234	-552.2944	-0.486913	430	0484	-524.3721	-0.056827	84
0235	-552.1348	-0.074047	94	0485	-524.2925	-0.312113	298
0236	-552.1215	-0.235379	234	0486	-524.2687	-0.173687	174

0237	-552.0896	-0.189172	184	0487	-524.0906	-0.218842	211
0238	-551.7679	-0.162063	162	0488	-524.0894	0.023542	33
0239	-551.716	-0.569998	474	0489	-523.9417	-0.571266	475
0240	-551.6425	-0.089932	106	0490	-523.8911	-0.220842	214
0241	-551.6257	0.068721	17	0491	-523.8814	-0.125895	134
0242	-551.2885	-0.257729	252	0492	-523.8773	-0.206892	201
0243	-551.1975	-0.17242	172	0493	-523.8413	-0.055536	82
0244	-550.916	-0.097537	109	0494	-523.7992	-0.235293	233
0245	-550.7216	-0.231649	226	0495	-523.7841	-0.341348	312
0246	-550.5222	-0.54991	462	0496	-523.7809	-0.227863	221
0247	-550.4511	-0.19774	191	0497	-523.7307	0.106267	6
0248	-550.2766	-0.238115	236	0498	-523.6931	-0.104251	114
0249	-550.1959	-0.618422	487	0499	-523.6926	-0.304974	293
0250	-550.0688	0.088823	10	0500	-523.6006	-0.228757	223

11 Appendix C: HyHEL-10 Fab-Lysozyme docking with Hex and STP

This appendix presents the results of ranking Hex docking orientations for the structure 3HFM. Table 11-1 shows the hex dockings sorted by STP scores. 200 Hex solutions are named from dock0001 to dock0200 (0001 being the one most preferred by Hex.

Table 11-1: Hex docking orientations and STP scores - HyHEL-10 Fab-Lysozyme Complex (Work Done for Masters Degree)

Dock Rank	T-Score	RMS	Dock Rank	T-Score	RMS	Dock Rank	T-Score	RMS
0058	0.203934	1.27	0120	-0.132131	59.42	0074	-0.213427	55.9
0012	0.155749	3.47	0126	-0.133534	55.15	0141	-0.21434	51.94
0118	0.151463	19.94	0155	-0.134965	27.61	0103	-0.21473	55.93
0114	0.1375	14.36	0156	-0.13506	56.97	0085	-0.216941	23.71
0144	0.113256	23.64	0093	-0.135263	54.92	0162	-0.217573	50.28
0117	0.076698	27.34	0167	-0.135507	26.59	0181	-0.217985	51.5
0197	0.043708	20.89	0168	-0.139606	58.61	0101	-0.220239	22.16
0135	0.036069	22.07	0052	-0.14021	55.01	0199	-0.221571	49.99
0175	0.034451	24.64	0018	-0.143084	53.88	0147	-0.22476	56.59
0069	0.03402	30.64	0033	-0.144934	55.33	0130	-0.22524	50.33
0165	0.01211	26.86	0064	-0.145366	56.31	0059	-0.230282	56.92
0127	0.012	27.43	0029	-0.146997	47.01	0151	-0.23036	51.73
0096	0.004395	25.51	0066	-0.147368	56.61	0070	-0.23073	40.09
0158	0.004391	28.39	0060	-0.148669	55.75	0034	-0.231795	51.24
0042	-0.007655	25.29	0124	-0.149004	55.96	0002	-0.23276	52.13
0026	-0.0085	27.84	0176	-0.149017	27.69	0194	-0.233704	42.4
0109	-0.009915	15.91	0172	-0.1494	52.34	0072	-0.234625	56.74
0079	-0.01401	50.96	0116	-0.149644	57.31	0099	-0.23544	47.78
0004	-0.017993	26.76	0160	-0.15289	51.85	0133	-0.235477	44.64
0146	-0.020717	27.89	0044	-0.153776	48.58	0166	-0.236373	54.03
0148	-0.020924	59.26	0081	-0.155925	57.55	0028	-0.237452	50.4
0183	-0.02283	55.71	0011	-0.156114	49.18	0091	-0.239383	53.52
0055	-0.032208	25.97	0189	-0.15641	57.08	0077	-0.240873	53.31
0100	-0.040586	26.08	0171	-0.156771	50.94	0152	-0.244841	52.85
0061	-0.042483	24.55	0041	-0.159774	53.11	0047	-0.248864	43.8
0123	-0.044773	53.49	0006	-0.160441	58.72	0132	-0.250122	48.05

0021	-0.047089	56.92	0071	-0.160777	52.63	0178	-0.250633	47.14
0186	-0.053359	27.42	0048	-0.161265	56.41	0136	-0.256906	44.31
0056	-0.053517	23.44	0137	-0.161382	27.05	0106	-0.257418	48.06
0078	-0.056675	28.57	0149	-0.164357	50.72	0025	-0.257749	41.15
0111	-0.061168	19.79	0180	-0.165481	52.7	0005	-0.257821	43.68
0020	-0.062677	24.51	0014	-0.16644	54.79	0038	-0.263874	49.53
0198	-0.064769	57.19	0046	-0.168763	27.01	0196	-0.268855	56.05
0054	-0.070559	25.65	0032	-0.17013	27.72	0015	-0.269427	48.04
0190	-0.073065	20.11	0088	-0.171131	59.33	0191	-0.270172	56.36
0031	-0.075772	54.63	0112	-0.171478	54.78	0082	-0.270749	57.14
0097	-0.077368	25.9	0016	-0.172506	55.45	0053	-0.272734	48.98
0023	-0.081722	56.73	0037	-0.17515	56.46	0007	-0.272893	56.12
0063	-0.082429	16.35	0045	-0.175514	45.97	0108	-0.273018	48.86
0128	-0.093623	53.48	0193	-0.176601	22.37	0057	-0.279691	51.21
0113	-0.096006	55.26	0134	-0.177119	56.41	0187	-0.283777	43.93
0105	-0.096453	57.64	0086	-0.179058	35.96	0185	-0.287212	47.88
0192	-0.099559	23.36	0008	-0.180698	32.02	0092	-0.288571	38.81
0139	-0.102364	27.58	0179	-0.182738	53.06	0049	-0.290694	56.8
0098	-0.103815	29.38	0062	-0.183372	46.44	0184	-0.294028	46.21
0075	-0.104625	19.51	0030	-0.183498	51.4	0022	-0.295272	45.09
0019	-0.105657	55.81	0087	-0.183946	27.88	0009	-0.29651	43.85
0001	-0.107519	30.36	0121	-0.18456	49.58	0080	-0.297273	55.62
0083	-0.10966	29.95	0068	-0.186104	56.66	0067	-0.298581	50.43
0076	-0.110702	58.5	0043	-0.186585	51.12	0125	-0.299336	46.04
0089	-0.111533	51.96	0157	-0.186635	51.02	0195	-0.299745	48.46
0150	-0.113516	21.02	0107	-0.186742	45.16	0174	-0.304154	54.28
0173	-0.113683	52.37	0188	-0.188179	52.56	0050	-0.304924	52.31
0065	-0.113762	20.48	0131	-0.191767	55.52	0095	-0.307184	50.16
0161	-0.115109	28.75	0094	-0.192947	49.72	0119	-0.321687	44.29
0170	-0.115281	28.5	0143	-0.19529	52.47	0073	-0.322583	57.35
0040	-0.117994	43.04	0051	-0.195324	49.82	0104	-0.324732	52.27
0102	-0.118624	45.94	0153	-0.196566	56.26	0024	-0.331766	47.05
0122	-0.121507	54.83	0017	-0.198049	42.42	0138	-0.337786	48.71
0084	-0.121565	52.26	0027	-0.200475	44.46	0035	-0.340198	54.72
0010	-0.122113	58.23	0169	-0.201014	47.32	0013	-0.353853	42.52
0145	-0.12659	48.18	0129	-0.202424	50.41	0164	-0.358889	31.49
0142	-0.12774	48.2	0115	-0.202539	57.9	0159	-0.359132	47.6
0154	-0.127886	40.73	0039	-0.203774	55.84	0003	-0.359755	46.16
0090	-0.128717	51.81	0140	-0.204971	55.02	0177	-0.384836	43.24
0036	-0.129502	51.09	0182	-0.207066	48.98	0200	-0.391792	46.08
0110	-0.130684	49.69	0163	-0.211807	50.85			

12 Appendix D: Programming Code of all the scripts Used in Chapter 6

Programming Code 12-1: The script that automates the creation of the ProPep dataset. All programs are found in *'/usr/people/wissam/Simon/perlScripts/wissam'*.

```
#include <stdio.h>
#include <string.h>
#include <stdlib.h>
int main(int argc, char **argv)
{
        system("./PDBextract_v3.pl");
        system("./which_chains_scibs.pl PDBextract_v3_chlist.log");
        system("wget ftp://ftp.wwpdb.org/pub/pdb/derived_data/NR/clusters90.txt");
        system("./clusterv2.pl clusters90.txt complex_nice3.log");
        system("./bestres.pl clusters90_outBIG.log");
        system("./interaction_extract.pl clusters90_outBIG_BEST.log");
        system("rm *.pdb *.hb2");
        system("./perl2sqlSDH.pl updatedList.log");
        system("cp *.log logs/");
        system("cp *.txt logs/");
        system("./torsion_out22.pl updatedList.log");
        system("./pdb_edit.pl updatedList.log");
        system("mv *.asa InteractionOutput/");
        system("mv *.rsa InteractionOutput/");
        system("cp logs/updatedList.log .");
        system("./runasa.pl updatedList.log");
        system("./runasa_atom.pl updatedList.log");
        system("./sql_update_amino acids.pl");
        system("wget ftp://ftp.wwpdb.org/pub/pdb/derived_data/NR/clusters50.txt");
        system("wget ftp://ftp.wwpdb.org/pub/pdb/derived_data/NR/clusters70.txt");
        system("./family_update.pl clusters50.txt Fam50");
        system("./family_update.pl clusters70.txt Fam70");
        system("./family_update.pl clusters90.txt Fam90");
        system("./update_BSA.pl");
}
```

Programming Code 12-2: Automation of the calculation of implicit database values.

```
#include <stdio.h>
#include <string.h>
#include <stdlib.h>

int main(int argc, char **argv)
```

```
{
        printf("1/8\n");
        fflush(NULL);
        system("echo \"UPdaTE PDB set Prob = 'OK' where Prob IS NULL\" | mysql -h
        cycfs -u wissam -p<password> ProPep08");

        printf("2/8\n");
        fflush(NULL);
        system("echo \"Create temporary table TEMP select PDBID, PepAtom,
                count(PDBID) as No from PepProVDW where Distance <3.8 Group By
                PDBID, PepAtom; Update PepAtom as P, TEMP set P.VDW_CN =
                TEMP.No where P.PDBID = TEMP.PDBID and P.AtomNo =
                TEMP.PepAtom\" | mysql -h cycfs -u wissam -p<password> ProPep08");
        system("echo \"UPdaTE PepAtom set VDW_CN = 0 where VDW_CN is NULL\" |
                mysql -h cycfs -u wissam -p<password> ProPep08");

        printf("3/8\n");
        fflush(NULL);
        system("echo \"Create temporary table TEMP select PDBID, PepAtom,
                count(PDBID) as No from PepProHB where Distance <3.8 Group By
                PDBID, PepAtom; Update PepAtom as P, TEMP set P.HB_CN = TEMP.No
                where P.PDBID = TEMP.PDBID and P.AtomNo = TEMP.PepAtom\" |
                mysql -h cycfs -u wissam -p<password> ProPep08");
        system("echo \"UPdaTE PepAtom set HB_CN = 0 where HB_CN is NULL\" | mysql
                -h cycfs -u wissam -p<password> ProPep08");

        printf("4/8\n");
        fflush(NULL);
        system("echo \"Create temporary table TEMP select PDBID, ProAtom,
                count(PDBID) as No from PepProVDW where Distance <3.8 Group By
                PDBID, ProAtom; Update ProAtom as P, TEMP set P.VDW_CN =
                TEMP.No where P.PDBID = TEMP.PDBID and P.AtomNo =
                TEMP.ProAtom\" | mysql -h cycfs -u wissam -p<password> ProPep08");
        system("echo \"UPdaTE ProAtom set VDW_CN = 0 where VDW_CN is NULL\" |
                mysql -h cycfs -u wissam -p<password> ProPep08");

        printf("5/8\n");
        fflush(NULL);
        system("echo \"Create temporary table TEMP select PDBID, ProAtom,
                count(PDBID) as No from PepProHB where Distance <3.8 Group By
                PDBID, ProAtom; Update ProAtom as P, TEMP set P.HB_CN = TEMP.No
                where P.PDBID = TEMP.PDBID and P.AtomNo = TEMP.ProAtom\" |
                mysql -h cycfs -u wissam -p<password> ProPep08");
        system("echo \"UPdaTE ProAtom set HB_CN = 0 where HB_CN is NULL\" | mysql
                -h cycfs -u wissam -p<password> ProPep08");

        printf("6/8\n");
        fflush(NULL);
        system("./Update_nodes.pl");

        printf("7/8\n");
        fflush(NULL);
```

```
        system("echo \"UPdaTE ProResidue, ProAtom Set ProResidue.Contact ='Y' where
                ProResidue.PDBID = ProAtom.PDBID AND ProResidue.ResidueNo =
                ProAtom.ResidueNo AND ProAtom.VDW_CN>0\" | mysql -h cycfs -u
                wissam -p<password> ProPep08");
        system("echo \"UPdaTE ProResidue Set Contact ='N' where Contact IS NULL\" |
                mysql -h cycfs -u wissam -p<password> ProPep08");

        printf("8/8\n");
        fflush(NULL);
        system("echo \"UPdaTE PepResidue, PepAtom Set PepResidue.Contact ='Y' where
                PepResidue.PDBID = PepAtom.PDBID AND PepResidue.ResidueNo =
                PepAtom.ResidueNo AND PepAtom.VDW_CN>0\" | mysql -h cycfs -u
                wissam -p<password> ProPep08");
        system("echo \"UPdaTE PepResidue Set Contact ='N' where Contact IS NULL\" |
                mysql -h cycfs -u wissam -p<password> ProPep08");
}
```

Programming Code 12-3: Automation of the creation of the Amino acid Pictogram. This code produces the necessary input needed to draw the figures.

```
#include <stdio.h>
#include <string.h>
#include <stdlib.h>

int main(int argc, char **argv)
{
        printf("1/6\n");
        fflush(NULL);
        system("echo \"SELECT A.Node, count(*) AS Int_Nodes, SUM(A.VDW_CN) AS
                Total_VDW, AVG(A.VDW_CN) AS Avg_VdW, STD(A.VDW_CN) AS
                SD_VdW, SUM(A.HB_CN) AS Total_HB, AVG(A.HB_CN) AS Avg_HB,
                STD(A.HB_CN) AS SD_HB, SUM(A.ASA_Solo-A.ASA_Comp) AS
                Total_RSA, AVG(A.ASA_Solo-A.ASA_Comp) AS Avg_RSA,
                STD(A.ASA_Solo-A.ASA_Comp) AS SD_RSA FROM PepAtom A, PDB P
                WHERE A.VDW_CN > 0 AND A.PDBID = P.PDBID AND P.Prob = 'OK'
                GROUP BY A.Node\" | mysql -h cycfs -u wissam -p<password> ProPep08
>               tables/NodeAnalysisPepAll.txt");

        printf("2/6\n");
        fflush(NULL);
        system("echo \"SELECT A.Node, count(*) AS Int_Nodes, SUM(A.VDW_CN) AS
                Total_VDW, AVG(A.VDW_CN) AS Avg_VdW, STD(A.VDW_CN) AS
                SD_VdW, SUM(A.HB_CN) AS Total_HB, AVG(A.HB_CN) AS Avg_HB,
                STD(A.HB_CN) AS SD_HB, SUM(A.ASA_Solo-A.ASA_Comp) AS
                Total_RSA, AVG(A.ASA_Solo-A.ASA_Comp) AS Avg_RSA,
                STD(A.ASA_Solo-A.ASA_Comp) AS SD_RSA FROM ProAtom A, PDB P
                WHERE A.VDW_CN > 0 AND A.PDBID = P.PDBID AND P.Prob =
                'OK' GROUP BY A.Node\" | mysql -h cycfs -u wissam -p<password>
                ProPep08 > tables/NodeAnalysisProAll.txt");
```

```
        printf("3/6\n");
        fflush(NULL);
        system("echo \"SELECT A.Node,count(*) AS Int_Nodes, SUM(A.VDW_CN) AS
                Total_VDW, AVG(A.VDW_CN) AS Avg_VdW, STD(A.VDW_CN) AS
                SD_VdW, SUM(A.HB_CN) AS Total_HB, AVG(A.HB_CN) AS Avg_HB,
                STD(A.HB_CN) AS SD_HB, SUM(A.ASA_Solo-A.ASA_Comp) AS
                Total_RSA, AVG(A.ASA_Solo-A.ASA_Comp) AS Avg_RSA,
                STD(A.ASA_Solo-A.ASA_Comp) AS SD_RSA FROM PepAtom A, PDB
                P, PepResidue R WHERE A.VDW_CN > 0 AND A.PDBID = P.PDBID
                AND A.ResidueNo = R.ResidueNo AND A.PDBID = R.PDBID AND
                R.SStruc = '1' AND P.Prob = 'OK' GROUP BY A.Node;\"|mysql -h cycfs -u
                wissam -p<password> ProPep08 > tables/NodeAnalysisα.txt");

        printf("4/6\n");
        fflush(NULL);
        system("echo \"SELECT A.Node,count(*) AS Int_Nodes, SUM(A.VDW_CN) AS
                Total_VDW, AVG(A.VDW_CN) AS Avg_VdW, STD(A.VDW_CN) AS
                SD_VdW, SUM(A.HB_CN) AS Total_HB, AVG(A.HB_CN) AS Avg_HB,
                STD(A.HB_CN) AS SD_HB, SUM(A.ASA_Solo-A.ASA_Comp) AS
                Total_RSA, AVG(A.ASA_Solo-A.ASA_Comp) AS Avg_RSA,
                STD(A.ASA_Solo-A.ASA_Comp) AS SD_RSA FROM PepAtom A, PDB
                P, PepResidue R WHERE A.VDW_CN > 0 AND A.PDBID = P.PDBID
                AND A.ResidueNo = R.ResidueNo AND A.PDBID = R.PDBID AND
                R.SStruc = '3' AND P.Prob = 'OK' GROUP BY A.Node;\"|mysql -h cycfs -u
                wissam -p<password> ProPep08 > tables/NodeAnalysisβ.txt");

        printf("5/6\n");
        fflush(NULL);
        system("echo \"SELECT A.Node,count(*) AS Int_Nodes, SUM(A.VDW_CN) AS
                Total_VDW, AVG(A.VDW_CN) AS Avg_VdW, STD(A.VDW_CN) AS
                SD_VdW, SUM(A.HB_CN) AS Total_HB, AVG(A.HB_CN) AS Avg_HB,
                STD(A.HB_CN) AS SD_HB, SUM(A.ASA_Solo-A.ASA_Comp) AS
                Total_RSA, AVG(A.ASA_Solo-A.ASA_Comp) AS Avg_RSA,
                STD(A.ASA_Solo-A.ASA_Comp) AS SD_RSA FROM PepAtom A, PDB
                P, PepResidue R WHERE A.VDW_CN > 0 AND A.PDBID = P.PDBID
                AND A.ResidueNo = R.ResidueNo AND A.PDBID = R.PDBID AND
                R.SStruc = '5' AND P.Prob = 'OK' GROUP BY A.Node;\"|mysql -h cycfs -u
                wissam -p<password> ProPep08 > tables/NodeAnalysisOther.txt");

        printf("6/6\n");
        fflush(NULL);
        system("echo \"SELECT R.AminoAcid, count(*) FROM PepResidue R, PDB P
                WHERE R.PDBID = P.PDBID AND P.Prob = 'OK' AND R.Contact = 'Y'
                AND R.AAType = 'N' GROUP BY R.AminoAcid\"|mysql -h cycfs -u wissam -
                p<password> ProPep08 > tables/AminoAcidComposition.txt");
}
```

Published Article